CUK-H006890-28-P6151

The British Journal for the Philosophy of Science

28

Volume 39

Published for the
British Society for the Philosophy of Science
by Oxford University Press
1988

P6151

CONTENTS

P 6151

ARTICLES

Page

1 QUANTUM PROPENSITON THEORY: A TESTABLE RESOLUTION OF THE WAVE/PARTICLE DILEMMA
Nicholas Maxwell

51 RESEARCH PROGRAMMES AND EMPIRICAL RESULTS
F. M. Akeroyd

59 CHALMERS ON METHOD
Barry Gower

67 REAL DISPOSITIONS IN THE PHYSICAL WORLD
Ian J. Thompson

81 PHILOSOPHICAL OBJECTIONS TO THE KINETIC THEORY
John Nyhof

141 SOLVING THE PROBLEM OF INDUCTION USING A VALUES-BASED EPISTEMOLOGY
Brian Ellis

161 A THEORY OF PROBABILITY
T. V. Reeves

183 ABSOLUTISM AND RELATIONISM IN SPACE AND TIME: A FALSE DICHOTOMY
Ian Hinckfuss

193 WHAT IS A LAW OF NATURE? A HUMEAN ANSWER
Peter Urbach

211 EMPIRICAL TESTS ARE ONLY AUXILLIARY DEVICES
Tyrone Lai

225 REALISM AND THE COLLAPSE OF THE WAVE-PACKET
Henry Krips

233 QUANTUM PHYSICS AND THE IDENTITY OF INDISCERNIBLES
Steven French *and* Michael Redhead

247 ZEEMAN-GÖBEL TOPOLOGIES
Adrian Heathcote

277 THE PARTICIPANT IRREALIST AT LARGE IN THE LABORATORY
Ian Hacking

295 PLURALISTIC ONTOLOGY AND THEORY REDUCTION IN THE PHYSICAL SCIENCES
Fritz Rohrlich

313 METAPHYSICS, MSRP AND ECONOMICS
J. C. Glass *and* W. Johnson

331 REMARKS ON THE MODULARITY OF MIND
Benny Shannon

353 HYPOTHESIS TESTS AND CONFIDENCE INTERVALS IN THE SINGLE GASE
D. J. Johnstone

361 AZANDE LOGIC VERSUS WESTERN LOGIC?
Timm Triplett

367 POPPER ON INDUCTION
Andrew J. Swann

375 ON CAUCHY'S NOTION OF INFINITESIMAL
Nigel Cutland, Christoph Kessler, Ekkehard Kopp *and* David Ross

421 HATTIANGADI'S THEORY OF SCIENTIFIC PROBLEMS AND THE STRUCTURE OF STANDARD EPISTEMOLOGIES
Marco Giunti

441 THE DARWINIAN SYNTHESIS: A CRITIQUE OF THE ROSENBERG/WILLIAMS ARGUMENT
G. Van Balen
449 DIMENSIONS OF OBSERVABILITY
Peter Kosso
469 PHILOSOPHY OF SCIENCE: FROM JUSTIFICATION TO EXPLANATION
Aharon Kantorovich
495 STATISTICAL EXPLANATION, PROBABILITY, AND COUNTERACTING CONDITIONS
Thomas R. Grimes
505 ARE ALL FALSE THEORIES EQUALLY FALSE?
Thomas Mormann

DISCUSSIONS

111 ON THE ALLEGED IMPOSSIBILITY OF INDUCTIVE PROBABILITY
Ellery Eells
379 WAS DESCARTES A LIAR? DIAGONAL DOUBT DEFENDED
Peter Slezak
389 BARROW AND TIPLER ON THE ANTHROPIC PRINCIPLE VS. DIVINE DESIGN
William Lane Craig
397 A NOTE ON SCHRÖDINGER'S CAT AND THE UNEXPECTED HANGING PARADOX
Jack M. Holtzman
521 RESPONSE TO EHRING'S 'PAPINEAU ON CAUSAL ASYMMETRY'
D. Papineau
527 HOW NANCY CARTWRIGHT TELLS THE TRUTH
Allan Franklin
531 A GREEN PARROT IS JUST AS MUCH A RED HERRING AS A WHITE SHOE: A NOTE ON CONFIRMATION, BACKGROUND KNOWLEDGE AND THE LOGICO-PROBABILISTIC APPROACH
Steven French
537 DAVY REFUTED LAVOISIER NOT LAKATOS
Arthur Zucker

REVIEW ARTICLES

117 KEYNES AS A METHODOLOGIST
Donald Gillies
263 THE VALUE OF A FIXED METHODOLOGY
John Worrall
403 LEVIATHAN AND THE AIR-PUMP: HOBBES, BOYLE, AND THE EXPERIMENTAL LIFE
Richard C. Jennings
411 HOW EXPERIMENTS END
Allan Franklin
541 CANTORIAN SET THEORY AND LIMITATION OF SIZE
Joseph W. Dauben
551 OBSERVATION AND OBJECTIVITY
Paul K. Moser

REVIEWS

131 POLLOCK, JOHN: *Contemporary Theories of Knowledge.* Review by Paul K. Moser
415 EVANDRO AGAZZI (Ed): *Epistemologia: La filosofia della scienza in Italia nel '900.* Review by André Mercier

MISCELLANEOUS

March ANNOUNCEMENT: BRITISH SOCIETY FOR THE PHILOSOPHY OF SCIENCE ANNUAL CONFERENCE 1988
June ANNOUNCEMENT: BRITISH SOCIETY FOR THE PHILOSOPHY OF SCIENCE ANNUAL CONFERENCE 1988
June ANNOUNCEMENT: BSPS CONFERENCE ON "PHYSICAL INTERPRETATIONS OF RELATIVITY THEORY"
June ANNOUNCEMENT: CONFERENCE ON PHILOSOPHY AND COSMOLOGY
Sept. ANNOUNCEMENT: LAKATOS AWARD IN PHILOSOPHY OF SCIENCE
Dec. ANNOUNCEMENT: BSPS PROGRAMME FOR 1989
Dec. ANNOUNCEMENT: BSPS SPECIAL INTEREST GROUP

INDEX OF AUTHORS

ARTICLES

Akeroyd, F. M. Research Programmes and Empirical Results, 51

Cutland, N., Kessler, C., Kopp, E. *and* Ross, D. On Cauchy's Notion of Infinitesimal, 375

Ellis, B. Solving the Problem of Induction using a Values-Based Epistemology, 141

French, S. *and* Redhead, M. Quantum Physics and the Identity of Indiscernibles, 233

Giunti, M. Hattiangadi's Theory of Scientific Problems and the Structure of Standard Epistemologies, 421

Glass, J. C. *and* Johnson, W. Metaphysics, MSRP and Economics, 313

Gower, B. Chalmers on Method, 59

Grimes, T. R. Statistical Explanation, Probability, and Counteracting Conditions, 495

Hacking, I. The Participant Irrealist at Large in the Laboratory, 277

Heathcote, A. Zeeman-Göbel Topologies, 247

Hinckfuss, I. Absolutism and Relationism in Space and Time: A False Dichotomy, 183

Johnstone, D. J. Hypothesis Tests and Confidence Intervals in the Single Case, 353

Kantorovich, A. Philosophy of Science: From Justification to Explanation, 469

Kosso, P. Dimensions of Observability, 449

Krips, H. Realism and the Collapse of the Wave-Packet, 225

Lai, T. Empirical Tests are only Auxilliary Devices, 211

Maxwell, N. Quantum Propensiton Theory: A Testable Resolution of the Wave/Particle Dilemma, 1

Mormann, T. Are all False Theories Equally False? 505

Nyhof, J. Philosophical Objections to the Kinetic Theory, 81

Reeves, T. V. A Theory of Probability, 161

Rohrlich, F. Pluralistic Ontology and Theory Reduction in the Physical Sciences, 295

Shannon, B. Remarks on the Modularity of Mind, 331

Swann, A. J. Popper on Induction, 367

Thompson, I. J. Real Dispositions in the Physical World, 67

Triplett, T. Azande Logic versus Western Logic? 361

Urbach, P. What is a Law of Nature? A Humean Answer, 193

Van Balen, G. The Darwinian Synthesis: A Critique of the Rosenberg/Williams Argument, 441

REVIEW ARTICLES

Dauben, J. A. Cantorian Set Theory and Limitation of Size, 541

Franklin, A. How Experiments End, 411

Gillies, D. Keynes as a Methodologist, 117

Jennings, R. C. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life, 403

Moser, P. K. Observation and Objectivity, 551

Worrall, J. The Value of a Fixed Methodology, 263

REVIEWS

MERCIER, A.	Agazzi, E. (Ed.): *Epistemologia: La filosofia della scienza in Italia nel '900*, 415
MOSER, P. K.	Pollock, J.: *Contemporary Theories of Knowledge*, 131

DISCUSSIONS

CRAIG, W. L. Barrow and Tipler on the Anthropic Principle vs. Divine Design, 389
EELLS, E. On the Alleged Impossibility of Inductive Probability, 111
FRANKLIN, A. How Nancy Cartwright tells the Truth, 527
FRENCH, S. A Green Parrot is just as much a Red Herring as a White Shoe: A Note on Confirmation, Background Knowledge and the Logico-Probabilistic Approach, 531
HOLTZMAN, J. M. A Note on Schrödinger's Cat and the Unexpected Hanging Paradox, 397
PAPINEAU, D. Response to Ehring's 'Papineau on Causal Asymmetry', 521
SLEZAK, P. Was Descartes a Liar? Diagonal Doubt Defended, 379
ZUCKER, A. Davy Refuted Lavoisier not Lakatos, 537

Brit. J. Phil. Sci. 39 (1988), 277–294 *Printed in Great Britain*

50/, 05
3779

The Participant Irrealist At Large in the Laboratory

IAN HACKING

---

1 *The Official Story*
2 *Irrealism*
3 *Assays*
4 *The Creation of Phenomena*
5 *Identity*
6 *What Can You Do with TRH?*
7 *Is TRH Thyrotropin Releasing Hormone?*
8 *Realism*

Soon it will be time to write 'Ten Years after *Laboratory Life*'. The notable book of that title, by Bruno Latour and Steve Woolgar, was published in 1979 and the research for the book was completed exactly a decade ago. The print run was exhausted. The book has just been reissued by another publisher, always a good sign of abiding interest. And interesting it is: even its old subtitle was a manifesto in its own right: *The Social Construction of Scientific Facts*. Social constructionalist attitudes to the natural sciences have been thriving for several years, especially thanks to the energetic and iconoclastic inquiries of the 'Edinburgh School' with its self-styled 'strong programme' for the sociology of knowledge. But Latour and Woolgar stand out for two interconnected reasons. First, their version of constructionalism is firmly rooted in laboratory science. Secondly, it is adamantly anti-realist, in precisely the domain, experimental science, where temptations towards scientific realism seem to me very strong.

Philosophers of science have much debated realism/anti-realism in the past decade. Hardly any of us mention this book, although it is, in my opinion, perhaps the most powerful anti-realist tract. No philosopher better exemplifies the vice of silence about this book than myself. In 1983 my *Representing and Intervening* argued that nothing to do with theories (representing) helps settle the endless debates about scientific realism/anti-realism. But it urged that experimental science (intervening) strongly leads one to realism about the entities postulated by theories. My 'experimental argument for scientific realism about entities' at the end of the book relies on the way in which we use entities in order to achieve effects or to study further phenomena. We polarize electrons in order to investigate the weak neutral current. That, I held, makes a sturdy realism about electrons difficult to assail, even if we can be sceptical or anti-realist concerning the exact truth of the theories about the electrons that we so manifestly and in great detail actually employ.

Whatever be the merits or demerits of my position, it was shameful not to examine the one outstanding piece of work then available, that took laboratory science seriously, and argued the strongest anti-realist doctrine in existence. The present short study is an attempt to make amends. Certainly there has been much constructionalism since 1979. Certainly the two authors have moved on and provided valuable work on several new topics. Certainly numerous writings about the laboratory have appeared in the last few years (some such as Knorr-Cetina [1981] and Lynch [1985] was based on research concurrent with Latour and Woolgar). Certainly other perspectives have been brought to bear, for example, Garfinkel's ethnomethodology. But there is something grand about this pioneering study that deserves some retrospective reflection on its achievements.

In what follows I shall not be seen, by fans of Latour and Woolgar, to do justice to the book. On the other hand I shall not defend scientific realism about entities, let alone my own brand of that doctrine based on experiment. I shall single out some features of *Laboratory Life* that seem to me powerful, and largely ignore those that seem to me weak.[1] Thus I replace the vice of silence by the vices of moderation. I am not intent on advocating or refuting the thoughts of Latour and Woolgar, as a whole. I do want philosophers of science to take this book very seriously, and not (if they are anti-constructionalists) to lump it with dogma or policy that they detest. I want my colleagues to get away from their idle ambulatory pondering about realism in the contexts of crassly simplified theories, and to attend to some of the realities of the experimental life.

---

[1] Some things I ignore because I have nothing to add, e.g., the authors' observations about career structures and what they call 'cycles of credit'. But there are others that I ignore because my emphasis is so different. Thus they have an extraordinarily inscriptional attitude to laboratory science. 'Between scientists and chaos, there is nothing but a wall of archives, labels, protocol books, figures and papers'. 'Even insecure bureaucrats and compulsive nominalists are less obsessed by inscriptions than scientists' (p. 245). Virtually all apparatus is seen as a collection of inscriptional devices. I am afraid I regard this as a symptom of the now out-dated fascination with the sentence so characteristic of Paris intellectuals in the late sixties. The use of the word 'inscription' here is attributed to Derrida (p. 88, n. 2). It has equally been proposed as the best translation of 'enoncé' in Foucault's earlier *Archaeology of Knowledge*. Foucault, the writer more amenable to change and development and learning from past experience, dropped this whole mode.

I THE OFFICIAL STORY

In 1977 Roger Guillemin and Andrew Schally were jointly awarded the Nobel Prize in medicine for their determination of the chemical structure of thyrotropin releasing hormone (TRH), a product of the mammalian hypothalamus, and thought to be of great importance in human endocrinology. Crudely speaking, the thyroid gland runs our metabolism, our growth and our maturation. The hormone thyrotropin, a protein secreted by the pituitary gland, runs the thyroid gland, and TRH triggers the production of thyrotropin. (Thyrotropin is often called TSH: thyroid stimulating hormone. I shall keep the acronym TRH, but use the word thyrotropin throughout.)

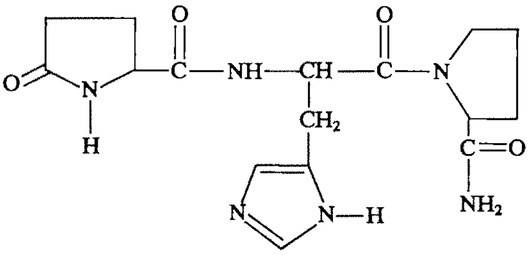
The research groups of Schally and Guillemin had been in fierce competition for some time. French-born Guillemin was a physiologist with a Ph.D. from l'Université de Montréal. Polish-born Schally had training in chemistry and a Ph.D. from McGill University across town. Guillemin established himself at the Baylor University College of Medicine in Houston, Texas. Schally worked as an assistant there for a short time before establishing himself at the Veterans' Administration hospital in New Orleans, and at Tulane University. Guillemin was cautious and painstaking. Schally is pictured as making lightning attacks on a problem. It is fairly easy to observe that each man feels *his* group made the discovery, a rivalry exacerbated by contrasts in training and technique. Guillemin's people wrote of Thyrotropin Releasing Factor, in part out of caution over the very concept 'hormone'. Schally's used the opposite terminology, and in that respect won the day. We have TRH, not TRF, although Latour, having worked with Guillemin, tends to the F-usage. (The substance is also called thyroliberin in e.g. the U.S.S.R., while the World Health Organization calls it protirelin.)

The strong interest in this work is due to a new twist in endocrinology. Instead of discovering the chemical structure of TRH by analyzing it, the two laboratories discovered it by synthesizing the substance. The problem of analysis is a consequence of the fact that TRH is so hard to come by. One hypothalamus secretes a minute amount of it, perhaps 20×10^{-9} grams.

Guillemin worked with pig brains. Five hundred tons of brains were sent down in refrigerated boxcars from the stockyards of Chicago to the laboratory in Texas. From this gigantic amount of pighead soup 1 mg of 80-percent TRH was extracted. Schally's people similarly started with 200 tons of sheep's brains for a comparably minute final product. There was too little of the hormone to provide for standard modes of chemical analysis. Even mass spectroscopy, used to obtain clues as to the molecules making up TRH, was problematic because the substance was not volatile.

After some false starts and changes of heart it became agreed that the hormone is a peptide, indeed a tri-peptide formed of three familiar amino acids in equal proportions. So it was possible to make each of the six possible chains of the three amino acids, and find out which behaved, under assay, like TRH. It did not quite work, but with very minor alterations to one end of the chain, the resulting substance duplicated the behavior of TRH. Thus was the chemical structure of the hormone discovered. The work won the Nobel Prize in medicine for two reasons. First, this hormone in the hypothalamus of the higher mammals controls thyrotropin that in turn controls the activity of the thyroid gland which in turn secretes hormones that control the metabolism, maturation and growth of the body. TRH looked like the key to human nature. Secondly, the idea of analysis by synthesis was deemed to be profoundly innovative.

There is now vastly more TRH—synthetic—in laboratories than there ever was in the entire universe before 1969. I shall describe in §6 below some of the empirical work done with TRH, mostly subsequent to the writing of *Laboratory Life*. But how could a couple of authors in their right minds ever imagine that these laboratories did not discover the structure of TRH? The labs chemically constructed a synthetic version of TRH, yes, but what has this to do with social construction? the brute fact is that TRH is:



pyroGlu———His———Pro–NH_2

2 IRREALISM

I should first discuss what *Laboratory Life* says about reality. 'Realism' is a notoriously difficult 'ism' to get a fix on, even when, as in recent years, philosophers of science have chosen to narrow the topic, speaking of scientific realism. One kind of scientific realism about theories holds that theories are true or false. A stronger version holds that whether they are true or false depends on whether they correspond to the world, or to the facts about the world. One kind of scientific realism about entities holds that the theoretical entities postulated by theories may exist. Both kinds of realism have an implied epistemological component: sometimes we have good reason to think that our theories are true. Or: some postulated unobservable entities, such as electrons, do exist. Theory realism and entity realism may be complementary but they are distinguishable. Thus Bertrand Russell once held that some scientific theories are true, but that the entities to which they refer do not exist. They are only logical constructions. Conversely, one may think that a good many theoretical entities exist, but that our theories about them are not true. (And think, moreover, that correspondence theories of truth are incoherent.) Thus Russell was a realist about theories and an anti-realist about entities, while I, for example, am a realist about entities but not about theories.

It is not to be expected that Latour and Woolgar, who have philosophical motivations but an anthropological approach, should neatly fit into the ephemeral philosophical categorization I have just described. There is no doubt, however, that they reject any picture that amounts to: 'There is a real world of facts; the aim of science is to describe this world; it does so by discovering facts'. They say, instead, that

> "reality" cannot be used to explain why a statement becomes a fact, since it is only after it has become a fact that the effect of reality is obtained. This is the case whether the reality effect is cast in terms of "objectivity" or "out thereness". It is *because* the controversy settles, that a statement splits into an entity and a statement about an entity; such a split never precedes the resolution of controversy (p. 180).

However, as they go on to say in a pair of sentences to which I add emphasis: '*We do not wish to say that facts do not exist or that there is no such thing as reality. In this simple sense our position is not relativist*' (pp. 180, 182). This is a quite different doctrine from that of the anti-realists who say that theories are only instruments to be used but not believed, or those who say that the aim of science is empirical adequacy, not truth. Latour and Woolgar report a world full of facts, but those facts are the historical product of 'microsociological processes.' There *is* a substance, TRH, secreted in minute amounts by the hypothalamus, and whose structure is that of a tripeptide, a string of three amino acids. That is a fact. But it became a fact.

There seems an air of trivial paradox here. Has not the hypothalamus of the higher vertebrates been secreting this substance ever since the animals came into being? Has it not always been a fact that this substance has a certain structure, a structure that became known in the laboratories of Texas and Louisiana? Latour and Woolgar do not say that something in the hypothalamus changed in 1969. But they think that what logicians would call the modality and tense structure of assertions of fact is misunderstood. Let *F* be a relatively timeless fact, say the fact that TRH has such and such a chemical structure. The official view would be: before 1969 one was not entitled to assert, categorically, that *F* is a fact, nor that *F* has always been a fact. But since then we know enough to be justified in asserting that *F* is a fact, and has always been so. Latour and Woolgar say no: Only after 1969 and a particular series of laboratory events, exchanges, and negotiations did *F* become a fact, and only after 1969 did it become true that *F* was always a fact. The grammar of our language prevents us from saying this. Our very grammar has conditioned us towards the timeless view of facts. But it is not only our grammar. In their own italics they write:

> *Our argument is not just that facts are socially constructed. We also wish to show that the process of construction involves the use of certain devices whereby all traces of production are made extremely difficult to detect* (p. 176).

Their picture of what is happening is this. In the beginning of research people don't have much sense of which statements are even in the cards, certainly not which are true or probable. But when a discovery is made, a statement begins, as they put it, to stabilize. We get a 'split entity', namely a statement about an object, and on the other hand the object that the statement is about: 'It is as if the original statement had projected a virtual image of itself which exists outside the statement' (p. 176). 'The past becomes inverted. TRF has been there all along, just waiting to be revealed for all to see' (p. 177).

This kind of anti-realism is not very familiar to philosophers who write in English. The established American philosopher whose views it most resembles is Nelson Goodman. Goodman calls himself an irrealist. His *Ways of Worldmaking* speaks of versions of the world. A dictionary will tell us that a version is an account or description from a particular point of view, especially as contrasted with another account. That suggests something neutral being described, or accounted for. Not for Goodman, for whom the version makes the world or world-fragment described. 'The differences between fitting a version to a world, a world to a version, and a version together or to other versions fade when the role of versions in making the worlds they fit is recognized.' [1978, p. 138]. In my opinion, Latour and Woolgar fit together with Goodman quite well: well enough to call them irrealists. This at least makes clear that what they maintain does not fall in the network of recent realism/anti-realism in the philosophy of science.

For example, the doctrine that Van Fraasen [1981] calls constructive empiricism is widely thought to be a controversial type of anti-realism about science. In fact it is extraordinarily conservative compared to the irrealism of Latour and Woolgar. For it holds that the aim of science is to achieve empirical adequacy, i.e., to predict what happens. It is an instrumental doctrine holding that the truth of theories is of no moment for their predictive empirical adequacy. I call this conservative because it does not call in question what it is to be empirically adequate, what it is to predict, what it is for a theory successfully to fit the phenomena, or what the phenomena *are*. All of these, according to Latour and Woolgar, are the product of negotiations among small groups of involved research workers. Whether or not a theory is empirically adequate to a body of phenomena is not something preordained in the theory and the phenomena, but rather something that becomes established (or not) as a scientific fact.

3 ASSAYS

One of *Laboratory Life's* more compelling considerations concerns assays. We have discovered the structure of TRH: such an assertion suggests that something (TRH) was given, and we investigated it. In fact there was no prior standard for recognizing TRH is a sample. Arguably it was a battle over assays—and hence a battle over what substances were under scrutiny—that helped choose the winners in the TRH competition. In 1960 many laboratories were asking about "releasing factors". These were narrowed to four, in Japan, Hungary, Texas and Louisiana. They had quite different ways of assaying for TRH, and the winners in the field won a first round by choice of assay.

Gold is the most familiar and perhaps the most ancient target for assaying. Assays were developed in a bootstrap fashion. Workers, craftsmen and consumers can 'just tell' gold. In certain forms it is a recognizable natural kind. It can be purified by melting. But it can also be adulterated. Long ago techniques evolved for testing the purity of gold.

The history of TRH is hardly like that. There were no primitive beginnings where we had a sample of recognizable if perhaps not pure TRH, and for which we elaborated assay techniques. 'Without a bioassay,' as Latour and Woolgar remark, 'a substance could not be said to exist' (p. 64). We did not begin with independent criteria of identity for thryotropin releasing factor (or hormone), and then develop techniques for isolating this independently identifiable substance. *The chosen assay system determined the criteria of identity for the substance.*

There are numerous kinds of assay. *Laboratory Life* gives some idea of the variety. Perhaps a better idea of the skills required in similar procedures, and the contrast between actual practice and official instructions, is provided by Lynch [1985] on pp. 69–76. A typical TRH bioassay might proceed like this, following Piva and Steiner [1972]. It is called the pituitary deletion method. We start with some material, say extracts from the hypothalamus. It is known in general to deplete the thyrotropin in the pituitary. We measure the extent of the 'releasing factors' in the material by seeing the extent to which it depletes the pituitary. This is done by injecting hypothalamic extracts into test animals. Then we compare the amount of thyrotropin left in their pituitary glands, and the amount that has been released into their bloodstream. These comparisons are made using a further bioassay, namely one for thyrotropin.

The number of variables here is staggering. For example, it is a practice to neutralize the hypothalamic material, because acids will deplete the pituitary on their own. Where do you inject material into the animal? You get a more direct effect if the stuff is put into a carotid artery than a vein. What animals do you use? Rats rather than mice, because 'it has been reported that the thyroid of mice is sensitive, in addition to' thyrotropin, to X, Y, Z and W.

What rats? How heavy? How old? Of what sex? About 70- or 80-day-old rats 'because it has been shown that pituitary [thyrotropin] content is maximal in 80-day-old rats.' Males, to avoid 'interference' by the female sexual cycle, and because males are thought to have more thyrotropin in the pituitary than females. When do you inject? At the same time of day, at least, for there are diurnal thyrotropin cycles. Moreover the rats should live in a standardized environment for at least a fortnight before treatment (14 hours of light alternating with 10 hours of dark). Do you anaesthetize the animals before injection and if so, with what (to be sure of no thyrotropin effect from the anaesthetic)? How long are they to live between injection and 'sacrifice'? Do you mix the serum taken from the dead animals and test it collectively? What do you do with bodies awaiting autopsy for the pituitary? And so on to the next stage, the thyrotropin assays using 'the classical McKenzie method as modified by Yamazaki *et al.*'

My account is taken from research shortly after TRH had been discovered and the issue of its structure settled. One can hardly overstate the evident amount of discretion even in the administration of one kind of assay. One would tell an analogous story in reporting radioimmunoassays. There, the idea is to create an antibody to TRH in the bloodstream of an animal, an antibody that is tagged with a radioactive tracer.

Latour reports that in the original research, Guillemin 'decided to set up a new assay based on the principle of the McKenzie assay [...], because a technician who had previously worked on [thyrotropin] came into the laboratory' (p. 125). It was partly this that settled Guillemin's group into looking for TRH rather than one of half a dozen other releasing factors of interest. Other kinds of assays were not checked. 'If you start checking that sort of thing, you never do anything,' or so Latour reports Guillemin as saying in interview. Guillemin also said that he could not understand how the Hungarian group had used their 'ludicrous assay, while anybody could have done what we did in 1961 and set up a true assay [...].'

Presumably the quotation is intended to be ironic. The concept of a 'true assay' is totally misleading. The assay that is chosen, settled on, creates the criteria of identity for the substance of interest. It creates its own truth. It is self-authenticating, for there is nothing else to authenticate it. How is an assay method settled upon? *Laboratory Life* describes the process as one of microsociology and negotiation. However we describe it, the metaphor of constructing criteria of identity for a substance is natural, if not inevitable.

4 THE CREATION OF PHENOMENA

Latour and Woolgar rightly draw attention to the standardized collection of equipment in the laboratory. The equipment has emerged from quite different sciences, and was, at one time, in those separate fields, as controversial as TRH in early 1969. It has congealed into technically reliable off-the-shelf products, just as now, in 1987, there is technically reliable synthetic TRH on the shelves of several hundred fundamental research laboratories around the world. Moreover, any physician can get some, and many have. Examples of such 1960s equipment would be the mass spectrometer or the nuclear magnetic resonance spectrometer. The difference from TRH is that the latter is cheap.

Our authors state that many phenomena would not exist without the standardized instruments. 'It is not simply that phenomena *depend on certain material instrumentation; rather the phenomena are thoroughly constituted by* the material setting of the laboratory. The artificial reality, which participants describe in terms of an objective entity, has in fact been constructed [...]' (p. 64). They recall Gaston Bachelard's discussion of these aspects of laboratory life, and Bachelard's description of it as *phénomenotechnique.*

These observations are of great importance, but the inferences to be drawn from them are not unequivocal. Take the contrast between 'artificial reality' and 'objective entity'. The word 'artificial' is ambiguous. In its first meaning, we have 'produced by man, not occurring naturally.' In the second, we have 'made in imitation of a natural product, especially as a substitute; not genuine; *artificial cream*'.

In the first and original sense of 'artificial' there is no contrast between being artificial and being an objective entity. Shoes are artificial, i.e., manufactured, but also objective entities. (Even shoes made of artificial leather are objective entities.) I have strongly urged a view consonant with Latour and Woolgar, but much distrusted by philosophers of science: one of the chief consequences of experiment is the creation of phenomena (see Hacking [1983] ch. 13). Virtually none of the phenomena by which we elaborate, articulate and test theories exists, in a pure state, in nature, before we create them. I find this not merely consistent with scientific realism (about entities and phenomena as opposed to theories) but positively an argument for it. Latour and Woolgar invite a conclusion opposed to mine. They are not deliberately equivocating on the two senses of 'artificial.' They are intending to say that facts are artificial in the sense of being manufactured (as opposed to phony). But a little of the attraction of their thesis may result from calling the phenomena created by experiment artificial, and inviting us to hear the implication of spurious, not 'objective,' not genuine, as artificial cream is to cream.

Nevertheless their recollection of Bachelard, together with Latour's sheer recognition of the fact of the creation of phenomena and of the stability of instrumentation, is important. It is one of the reasons why, in my opinion, this early work on the anthropology of the laboratory surpasses much that has been published subsequently.

It is chiefly philosophers with no experience of a laboratory that can doubt the doctrine of the creation of phenomena. One of the best passages in Latour's self-description is a reference to his clumsiness and incompetence as a technician. Honestly reporting everything, including his blushing and his loss of a day's work, he concludes, 'It is necessary to be a technician, and an incompetent one at that, in order fully to appreciate the practical miracle (in Bolzmann's sense of the word) which gives rise to a standard curve' (p. 245).

5 IDENTITY

What is identical to what? There is a synthetic chemical, a tripeptide, a string of three amino acids written pyrtoGlu-His-Pro-NH_2. I shall call this *P* for short. It is conjectured that there is a factor in the hypothalamus that releases a hormone, thyrotropin, from the pituitary. I shall call this conjectured (and by implication) unique substance *H* for short. We commonly call *H* TRH, and if we bother to make distinctions, call *P* synthetic TRH. In practice hardly anyone bothers to make a distinction. *P* and *H* are identical: they are TRH. *This is the scientific fact under investigation by Latour and Woolgar:* $P = H$, a fact summed up in a name, TRH.

There are no problems about the criteria of identity for the synthetic *P*. We make it in large quantities, and if we think the manufacturer has made a mistake or is cutting it with sodium bicarbonate we can routinely check by that array of instruments mention in §4. In §3, on assays, we found a good deal of ground for scepticism—or at any rate for talk of construction—concerning the very identity of *H*, the hormone in our hypothalamus. But suppose we have constructed the criteria of identity for *H*, namely what is measured by a certain family of assays. Then we go on to ask, is *H P*?

Here there is a practical problem. In 1969 there was (thought to be) little *H* in the whole universe, maybe 100 g, or perhaps the equivalent of a 1/4-pound stick of butter. Commonly, if we were interested in *X* and solved its chemical constitution by synthesizing *Y*, then we would perform all sorts of routine tests to see whether *X* is identical to *Y*. But here, there simply isn't any *X* available. Having solved the problem in an attractive way, no one is going to confirm it by brute force, for that would involve getting another 500 tons of pig brains from Chicago. No one would be crazy enough to suggest that to the National Institutes of Health. So what happened was that *H* simply became *P*. In 1981,

> just twelve years ago, the field of endocrinology received one of its most significant advances when the structure of thyrotropin-releasing hormone, TRH, was announced. The close of the 1960's saw the culmination of more than fifteen years' effort by many laboratories when two research groups under R. Guillemin and A. V. Schally isolated and independently identified an identical structure for TRH from ovine and porcine hypothalami. (Griffiths and Bennett [1983], p.1).

One could describe this a little differently. Guillemin and Schally did synthesize the same substance, the 'identical structure' *P*. But they did not exactly identify *H as P*, in the way in which one says, I can identify the man in the green coat as the man who stole my watch. Rather they identified *H with P*. They made *H* and *P* identical. As Latour and Woolgar would say, the identity of *H* and *P* became a fact without the normal follow-up scrutiny that would accompany such a proposal.

Why not? Sheer impracticality, in part. There were also certain interests. The identity of a hormone in the hypothalamus may seem a model question for disinterested research. Not so. Hormones are very hot commercial products. At least in this case they would affect growth. That is not likely to be commercially very valuable. But they also probably affect sexual maturation, and that is big money. Moreover, there was much talk that TRH might affect mental function and be relevant to depression. That is really big money. Dr R. O. Studer [1972] of the Department of Chemical Research at F. Hoffman-La Roche & Co., Ltd, Basel, recalls the very day when Guillemin's tentative results were broadcast. 'It was at this moment, in January, 1969, when the existing loose contact between Roche and Dr. Guillemin was, of course, immediately tightened and the 2 groups acted as a single team' (p. 7). Of course.

The issue of *H* in the hypothalamus was closed off. *H* is *P*, namely TRH. But suppose it became suspected that there is TRH in parts of the body other than the hypothalamus? Latour and Woolgar would probably have predicted utterances like the following, from Jackson [1983a], p. 7:

> The presence of substantial quantities of TRH in the extrahypothalamic nervous system supports the view that TRH has a neuronal function in these location. However, Youngblood et al (61) have reported that the immunoreactive (IR)-TRH in extrahypothalamic regions is not identical to the tripeptide amide. This issue was of much importance since a neurocrine hypothesis to explain the effects of TRH in these regions would be undermined if the IR-TRH located there was not authentic. However, extracts of hypothalamus, extra-hypothalamic brain and spinal cord when subjected to high-performance liquid chromatography (HPLC) following purification on affinity chromatography (26) have shown identity of the tissue TRH from all sites with synthetic TRH (Fig. 3). Other groups (33, 54) have reported similar findings, and it has been suggested that Youngblood's discordant results reflect poor tissue extraction and/or interference in the TRH assay by their thin-layer gels (35).

6 WHAT CAN YOU DO WITH TRH?

Nearly 600 papers on TRH are listed in the 1986 *Biological Abstracts. Index Medicus* and *Chemical Abstracts* have fewer and many overlap with the biology selection. Say that at least 1,000 research papers about the substance are published every year.

The research falls under a number of recognizable categories. There is much assay improvement. This in turn leads to questions about where in the bodies of living organisms one can find TRH, how one can stimulate or simulate its production, and what increases in production do to other chemicals in the body.

There is also the development and investigation of analogs to TRH, a feature already noted in *Laboratory Life* (p. 61). As Latour and Woolgar would have predicted, there has emerged a standardized procedure for human subjects, the TRH test. There are a number of diseases for which this is to be expected, namely diseases connected with the thyroid. It is now a standard diagnostic procedure for patients with hyperthyroidism and euthyroidism. About half the patients tested experience 'nausea, flushing, light headedness, a metallic taste in the mouth and a deep urethral burning sensation,' as Kerr and Alexander [1984] report under the heading 'Is the TRH Test Really Necessary?' The authors suggest one gets equally good diagnostic results with a less highbrow test involving thyrotropin. Kirkegaard [1985] retorts that there remains a great deal of routine clinical work done with the standard physician's TRH test.

Then there is a lot of purely empirical investigation, in which TRH is administered to patients or other deviants that medicine and society are unable to help in any systematic way. Women with a history of attempted suicide are given massive doses of TRH to see what happens. (Linkowsky [1983], [1984]). Studies of TRH levels in reformed alcoholics are conducted. (Loosen *et al.* [1983]). The TRH test is used to see if it can sort out people with 'borderline personality disorders.' (Barbut *et al.* [1983]). The expectation is that one would have a routine biological procedure to see who should be prescribed mood-altering drugs. Winokur *et al.* [1982] study 'Improvements in Ratings of Tension after TRH Administration in Healthy Women'. An anorexic is noted to change her behavior after regular TRH injections. One thinks of Yarborough *et al.* [1985] who investigate 'The Therapeutic Potential of Thyrotropin Releasing Hormone (TRH) in Alzheimer's Disease (AD).' In short you shoot up people with the new chemical and see what happens. It is to be remembered that in the original evaluations it was expected that the human body normally had about 20×10^{-9} grams of TRH in it. Some of the subjects are now being routinely injected with more TRH than was thought to exist in all the human bodies of New York City.

Animals can be given much larger doses. See what happens! There is 'Head Turning Induced by Unilateral Intracaudate Thyrotropin-releasing Hormone (TRH) Injection in the Cat.' (Malovin *et al.* [1982]). There are 'Effects of Acute and Long-term Treatment with Thyrotropin-releasing Hormone on Locomotor Activity and Jumping Behavior in Mice.' (Ishujima *et al.* [1986]). Or you can keep a male frog away from females for a week, and see if it ejaculates on being injected with TRH (you kill it and analyze its urine to investigate that).

I choose somewhat garish examples partly just because they are there, and partly because they could usefully serve as a supplement to Latour and Woolgar's line of argument. It is evident that two things have come into being: first, large amounts of this stuff, TRH. Secondly, we now have a large number of testing, assaying and identifying procedures which have been transferred, in some cases, into off-the-shelf devices that can be used by semi-skilled workers. It is just this trajectory which (one infers) Latour and Woolgar would have predicted.

Perhaps they would not have predicted another field of research, namely the extension of TRH studies to lower orders on the evolutionary scale. If you starve an alligator for several months—only a modest enforcement of the alligator's inclination to eat little at certain times—and then cut it up, one finds (Jackson [1983b]) a great deal of TRH in the stomach of the reptile. Naturally this in itself suggests very interesting questions for the theory of evolution. How is it that chemicals in the stomachs of alligators are also critical to the human hypothalamus? We seem to have come some distance from the brain of the higher mammals. This leads me to my next question.

7 IS TRH THYROTROPIN RELEASING HORMONE?

Isn't it a mere tautology to answer in the affirmative? Not quite. The motivation of fifteen years of research, with its happy conclusion in 1969, was clear. There was the overall model of the factors that release hormones. There was an interest in discovering the chemical nature of *any* of these factors. Each was conceptually defined as a factor that releases a hormone. Synthetic TRH won a Nobel Prize partly for the ingenuity of the work, but chiefly because it would lead to a deeper understanding of the human body's control system. It might teach us much about human nature itself.

I have just described experiments on starved alligator stomachs. What has that got to do with the hypothalamic control system of the higher vertebrates?

The standardized assaying procedures—always evolving, of course—lead to the following remarkable conclusion. TRH is to be found in many quarters of the human body, or, at any rate in the more accessible bodies of rodents and rabbits. It is all over the central nervous system. Cooper *et al.* [1978] suggest that here it has a role as a neurotransmitter. It is reported in the islands of Langerhans in the pancreas, and in the gastrointestinal tract. Body fluids from blood to urine have lots of TRH.

I've mentioned TRH in the stomachs of starved alligators. Its most dense location in any organism may be the skin of the frog (and also the South African clawed toad, fed on a diet of chopped beef livers as is shown in a sequence of papers in Griffiths and Bennett [1983], pp. 191–228. Eels are spectacular in this respect. (Youngs *et al.* [1985]). Although artificial TRH does, in laboratory conditions, stimulate thyrotropin production in mammals and birds, it has no effect of the thyrotropin levels of amphibia such as frogs, toads and eels.

Fish too contain TRH. In fact it was found by Coller *et al.* [1982], p. 57, that for 'the goldfish and the trout the olfactory tubercle contained a higher concentration of TRH than did the hypothalamus.' TRH certainly does not increase production of thyrotropin in fish. There is some evidence for a thyrotropin *inhibiting* factor (TIF). It has been proposed by Bromage [1975] that TRH is none other than TIF for fish.

Now then, is TRH Thyrotropin releasing hormone?

8 REALISM

In one important respect Bruno Latour was very judicious or just plain lucky. A couple of years after he began work on his anthropology of the laboratory, his specimen of experimental work won a Nobel prize. He does not emphasize this in the book. But it does mean that the reader cannot say, 'Well, that's how it was in this crazy laboratory where Latour set up shop, but that's not real science.' He chose something with the official accolade of best science. Indeed Schally and Guillemin are beggars for punishment. Between them they have picked up honorary doctorates at nineteen universities in six different countries. *Laboratory Life's* constructionalist thesis, that Schally and Guillemin did not make a 'discovery', is not discussing men of straw.

Latour and Woolgar cannot be faulted for failing to discuss the best science around. But clearly they hold their story to be typical. But notice that the very features that make the TRH example attractive to a conservative reader such as myself, are not typical. It all boils down to that tiny amount of organic TRH extracted from pig and sheep brains. The reason that the assay looked so remarkably like a construction, and the reason that the synthetic stuff in effect defined the criteria of identity for TRH, is that there is so little of the natural stuff to go round. These features are lacking, in my opinion, in what must be the classic piece of discovery in organic chemistry, the Kekulé formula for the benzene ring (and/or its quantum mechanical re-interpretation). I do not deny that one can tell a constructionalist story for the benzene ring—just that it is not so compelling.

Let us suppose, however, that we generalize from the example of Woolgar and Latour, or at any rate take it to be sufficiently important in itself that it cannot be dismissed as mere anomaly. Does it compel us to irrealism or some other brand of anti-realism? I think not. It is usually possible to give a realist version of an anti-realist doctrine, and vice-versa. There is a change in tone, which will make some advocate one version over the other, but there may be substantial agreement all the same. There is, I think, a realist version of the kind of irrealism espoused by Latour and Woolgar.

When I speak of realism, I do not mean realism with all its philosophical accretions. For example, many scientific realists argue for realism in terms of our converging on the one great truth, or at least converging on something worth calling *the* truth. There is nothing, however, in the realist attitude that demands that there is a humanly expressible, uniquely richest statement of how the world is. Realism does not demand humanly accessible unity as an ideal, any more than theism demands monotheism. Theism implies monotheism only if unity is a perfection. It is astonishing that there are philosophers who honor Frege (existence and numbers are properties of properties, not properties of entities) and still hold that realism demands convergence on the one true and knowable body of sentences.

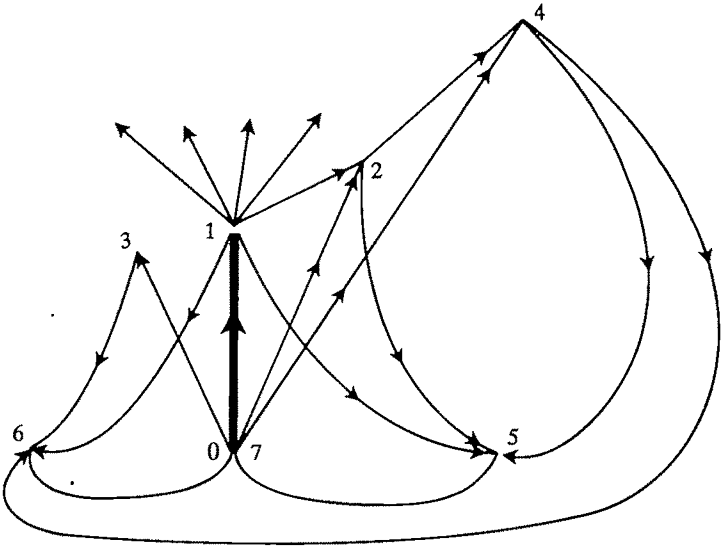
A realist version of Latour and Woolgar would have no truck with an ultimate end-product of science (aside from the contingent one of wherever we get to before we quit). I shall express a realist version using a correspondence theory of truth, simply because that is how most realists think. My own entity and experimental realism can also be written in a parallel way, but here it is better not to be tendentious.

The correspondence-theory-of-truth-realist should suppose that the universe is vastly too complex for people to describe all of it or to produce a unitary theory of it. We can instead imagine an endless array of world descriptions which are largely independent of each other. Each description is true of the world, and if you like corresponds to some aspects of the world's structure.

Each established description closes off other possible descriptions, not because they are mutually inconsistent, but because having reached a stage in one version, a benchmark, a whole new range of inquiries is opened up. Each description is expandable in many ways. From some sort of God's-eye view the process of our knowledge is like that of choosing forks in branching possible descriptions. Thus at stage 0 in 1960 we could imagine God seeing the following possibilities for our development, among many others.

1. There is a thyrotropin releasing factor of unique interest to us, and its structure is pyroGlu-His-Pro-NH_2.
2. There is a substance, namely pyro-etc., found throughout the zoological orders, and which serves a large number of control functions, no one of which is primary.
3. There is a thyrotropin releasing factor whose structure is X, and of unique interest to us, but substantially different from pyro-etc. Pyro-etc. never enters our field of interest, just as the five abandoned tripeptides originally synthesized along with pyro-etc. are of no interest. They are just structures that we could synthesize *en masse* but don't want to.
4. Dozens of thyrotropin releasing factors in the hypothalamus engage our attention. The endocrine system exhibits high redundancy, with many chemicals able to do many overlapping jobs. Moreover, all these factors are found elsewhere in the several zoological orders, and all have lots of functions.
5. The master-slave dominance model (factor *X* triggers hormone *Y* to produce activity Z) is all wrong. Science ceases to betray the masculine gender of which Keller [1985] has made us all conscious.
6. We all become Zen Buddhists.

7. Nothing happens. We stay at stage 0, the state of knowledge of 1960, at least so far as concerns releasing factors. In due course these research programs are abandoned.



On the diagram, the bold double line indicates our actual passage from 1960 to 1970, which is stage 1. The single thin lines represent *possible* transitions. From stage 1 we might move in many directions, including giving up the thyrotropin releasing emphasis in TRH and landing at stage 2 and even going to 4. But we might have got to 2 or 4 at once. Either in 1960 or in 1970 we could conceivably have made the revolutionary downlooping moves to 5 or 6. If we had got to 3, we would not get to 1 or 2, except by a revolutionary looping move. It would be very unlikely that we would get to 1 from 5 or 6. Finally, state 0 (1960) could have persisted. No progress is made, and the search for releasing factors comes to an end, out of frustration, boredom, career failure, and lack of funding. Conceivably this 7 might incline us to 5 or 6.

The fact that we have got to stage 1, namely TRH, opens up an enormous range of investigations of which only a few trifling examples were given in §6.

From this point of view, we really did discover a fact when we got to stage 1. But it was not the only fact we might have discovered. With 1, rather than 0, as a launching pad, our knowledge is very likely to take off in many directions that were unthinkable or impracticable from 0, and which would be so from 2, let alone 5. The world has within it all the facts of 0, 1, 2, 3, 4 and a whole range of other facts that might be engendered by 5. There is no call for irrealism on the basis of the observations by Latour and Woolgar. Their microsociology is used, in this realism, not to explain the construction of a scientific fact, but to explain the selection of fact 1 to enter our stream of knowledge. That is to say, it explains why we discovered the structure of TRH, rather than some other facts.

Work such as that by Latour and Woolgar attracts attention by its irrealism or its constructivism. I can have no objection to that, having hung a very different book about experiment on a realist coat-hanger. But the interest of work such as *Laboratory Life* is not its irrealism, no more than the interest of my book is its experimental realism. What is great about Latour and Woolgar is their detailed and articulate attention to experiment, a proper study for philosophers of science.

Institute for History and Philosophy of Science and Technology
University of Toronto

REFERENCES

BARBUT, J. C. *et al.* [1983]: 'The TRH Test in Patients with Borderline Personality Disorder.' *Psychiatry Research*, 9, pp. 107–113.

BROMAGE, N. R. [1975]: 'Thyrotropin Inhibiting Factor in Fish.' *General Comparative Endocrinology*, 25, pp. 292–7.

COLLER, J. *et al.* [1982]: 'The Central Nervous System,' in G. H. Gass and H. M. Kaplan (eds.), *Handbook of Endocrinology*, New York: CRC Press.

COOPER, J. R. *et al.* (eds.) [1978]: *The Biochemical Basis of Neuropharmacology*, 3rd. ed. New York: Oxford University Press.

EXTEIN, I. *et al.* [1984]: 'The TRH Test in Affective Disorders: Experience in a Private Clinical Setting,' *Psychosomatics*, 25, pp. 379–386.

GOODMAN, NELSON [1978]: *Ways of Worldmaking*, Indianapolis and Cambridge, Mass.: Hackett.

GRIFFITHS, E. C. and BENNETT, G. W. (eds.) [1983]: *Thyrotropin-Releasing Hormone*, New York: Raven.

HACKING, I. [1983]: *Representing and Intervening*, Cambridge: Cambridge University Press.

JACKSON, IVOR M. D. [1983a]: 'Thyrotropin-releasing Hormone (TRH) Distribution in Mammalian Species and its Functional Significance,' in Griffiths and Bennett 1983, pp. 6–19.

JACKSON, IVOR M. D. [1983b]: 'Changes in TRH Levels in Alligator Stomachs During Fasting.' *Peptides*, 9, pp. 311–314.

KELLER, EVELYN FOX [1985]: *Gender and Science*, New Haven: Yale University Press.

KERR, D. J. and ALEXANDER, W. D. [1984]: 'Is the TRH Test Really Necessary?' *Lancet*, (ii), p. 1162.

KIRKEGAARD, C. *et al.* [1985]: 'Is the TRH Test Necessary?' *Lancet*, (i), p. 222.

KNORR-CETINA, K. [1981]: *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*, Oxford and New York: Pergamon.

LATOUR, B. and WOOLGAR, S. [1979]: *Laboratory Life: The Social Construction of Scientific Facts*, Beverly Hills and London: Sage.

LATOUR, B. and WOOLGAR, S. [1986]: *Laboratory Life: The Construction of Scientific Facts*, 2nd. ed. with altered title, index and additional references and a postscript, but otherwise identical in all respects, including pagination, to Latour and Woolgar [1979]. Princeton: Princeton University Press.

LINKOWSKI, P. *et al.* [1983]: 'Thyrotropin Response to Thyrostimulin in Affectively Ill Women in Relation to Suicidal Behaviour.' *British Journal of Psychology*, **143**, pp. 401–405.

LINKOWSKI, P. *et al.* [1984]: 'Violent Suicidal Behavior and the Thyrotropin-releasing Hormone Thyroid-stimulating-test: A Clinical Outcome Study.' *Neuropsychobiology*, **12**, pp. 19–22.

LOOSEN, P. T. *et al.* [1983]: 'TRH in Abstinent Alcoholic Men.' *American Journal of Psychology*, **140**, pp. 1145–1149.

LYNCH, M. [1985]: *Art and Artifact in Laboratory Research: A Study of Shop Work and Shop Talk in a Research Laboratory*, London: Routledge Kegan Paul.

MALVOLIN, F. *et al.* [1982]: 'Head Turning Induced by Unilateral Intra-caudate Thyrotropin-releasing Factor in the Cat.' *European Journal of Pharmacology*, **81**, pp. 559–567.

PIVA, F. and STEINER, H. [1972]: 'Bioassay and toxicology of TRH.' *Frontiers of Hormone Research*, **1**, pp. 11–21.

STUDER, R. O. [1972]: 'Chemistry of TRH.' *Ibid.* pp. 3–14.

USHIJIMA, I. *et al.* [1986]: 'Effects of Acute and Long-term Treatment with Thyrotropin Releasing Hormone on Locomotor Activity and Jumping Behavior in Mice.' *Pharmacology, Biochemistry and Behavior*, **24**, pp. 1423–1428.

VAN FRAASSEN, BAS C. [1981]: *The Image of Science*, Oxford: Clarendon Press.

WINOKUR, A. *et al.* [1982]: 'Improvements in Ratings of Tension after TRH Administration in Healthy Women.' *Psychoneuroendocrinology*, **7**, pp. 239–244.

YARBROUGH, G. G. *et al.* [1985]: 'The Therapeutic Potential of Thyrotropon Releasing Hormone (TRH) in Alzheimer's Disease (AD).' *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, **9**, pp. 285–289.

YOUNGS, L. J. *et al.* [1985]: 'Thyrotropin-releasing Hormone in the Lamprey Central Nervous System.' *Brain Research*, **338**, pp. 177–180.

Brit. J. Phil. Sci. 39 (1988), 295–312 *Printed in Great Britain*

Pluralistic Ontology and Theory Reduction in the Physical Sciences

FRITZ ROHRLICH

ABSTRACT

It is demonstrated that the reduction of a physical theory S to another one, T, in the sense that S can be derived from T holds in general only for the mathematical framework. The interpretation of S and the associated central terms cannot all be derived from those of T because of the qualitative differences between the cognitive levels of S and T. Their cognitively autonomous status leads to an epistemic as well as an ontological pluralism. This pluralism is consistent with the unity of nature in the sense of a substantive monism.

1 *The Problem*
2 *The Argument for Substantive Monism from Cosmic Evolution*
3 *What is Meant by a Scientific Theory?*
4 *Mathematical Interlude*
5 *Views on Reduction*
6 *Reduction of the Mathematical Component*
7 *The Language Component*
8 *Pluralistic Epistemology*
9 *Pluralistic Ontology*

I THE PROBLEM

The problem to be stated exists in all the sciences but is most explicitly seen in the physical sciences which will therefore be used for illustrations.

Progress has led to a complex net of scientific theories spread over the many different branches of science. These theories range from the least fundamental to the most fundamental, and the more fundamental theory is considered to be superior and thus 'supersedes' the less fundamental one. Thus, chemistry tells us that a piece of wood is 'really' a complicated arrangement of many kinds of molecules bound together; atomic physics tells us that molecules are 'really' various atoms held together by interatomic forces; particle theory tells us that

---

* I am grateful to several people who provided helpful and constructive comments on an earlier version of this paper: Jim Cushing, Larry Hardin, Tom Nickles; Abner Shimony, and Ron Yoshida.

atoms are 'really' elementary particles in interaction, and so on. Eventually, one comes to the most fundamental theory of the day. We have a procession of 'models of the world' passing before us as the theories which demand them are passing, and we have barely accepted one when it is superseded by the 'real' one which is soon superseded by yet another 'real' one. What are we to make of this?

First of all, we must accept the fact that there is no reason to expect that the most fundamental theory of today is the ultimate one. This despite the pundits in every generation who proclaim the end of a particular science such as physics. It is not even clear that it is meaningful to speak of an ultimate fundamental theory, but that is not the issue here.

The problem here is not the successive compositeness of the objects in question until one arrives at ultimate primitive and noncomposite entities (quantum fields?) which are presumably very few in number. The problem is rather the ontology of each and every scientific theory that is superseded but not discarded. Newton's corpuscular theory of light was superseded (by the wave theory) *and* discarded, but his mechanics was superseded (by relativistic mechanics) and survived. We shall not deal with the discarded theories here. It is the nondiscarded ones that cause the problem: the old and the new theories have different ontologies; they are even claimed to be logically incompatible with one another or at least 'incommensurate' (Feyerabend [1962], Kuhn [1970]). Yet, *both* theories have certain claims to reality.

One common view is to *replace* the old theory by the new one and to pay little attention to its somewhat mysterious survival. Consistency then requires us to wait for the ultimate theory in order to be sure to have the correct ontology of nature and to accept all others as temporary and not to take them too seriously. I shall call this view of replacing ontologies by successively better and better ones aiming to reach the 'only true ontology' (of the ultimate theory) 'ontological monism'.

What alternative is there? How can the ontologies of hierarchically related scientific theories be mutually consistent? Can a 'pluralistic ontology' be consistent?

A well-known and fairly trivial example are the two levels of the 'furniture of the world': tables and chairs on one level, atoms and molecules on the other. But let us consider a more interesting example, one that involves two scientific theories.

Newtonian gravitation theory explains the motion of stars and planets by means of a *universal gravitational force* which is given ontological status in that theory. Einstein's gravitation theory (general relativity) is said to supersede Newton's because it explains the same phenomena and more. But it does so by a different ontology. The theory involves different central terms such as the *curvature of spacetime* which belongs to the ontology of that superseding theory. Newton's universal gravitational force is not a central term of

Einstein's theory; it emerges only in a suitably chosen limit under suitable assumptions. Nor is the notion of curved spacetime contained in Newton's theory. One way to resolve this difficulty is to consider Newton's theory simply wrong and to embrace Einstein's as the correct explanation of gravitational phenomena: Newton's ontology is to be *replaced* by Einstein's.

The replacement view encounters several difficulties.

(1) Einstein's gravitation theory is not the ultimate theory either. It will be superseded by yet another theory presently called 'quantum gravity theory' which will be a synthesis of general relativity and quantum mechanics (ignored in Einstein's theory). The development of quantum gravity is presently in the forefront of gravitation research. One already foresees a theory with a new ontology superseding Einstein's.

(2) Replacing Newtonian gravitation theory is contrary to scientific practice. Physicists and astronomers use the theory all the time even today, more than seventy years after the invention of general relativity. And beyond that, the theory is presently being developed further by means of improved computational techniques for various applications such as space flight.

(3) Newton's theory offers a model for the explanation of gravitational phenomena which (within very wide but well specified limits) is (a) empirically adequate, (b) conceptually much simpler than Einstein's theory, and (c) pragmatically preferable.

In the present paper, I shall argue against the monistic ontological view, *i.e.* against ontology replacement (and in fact against theory replacement) in cases such as the above (to be made specific below). I shall advocate a *pluralistic ontology* of science. To this end I must investigate *scientific theory reduction* which seems to be in tension with a pluralistic ontology: does not reduction imply that one ontology follows from another?

I shall further argue that it is our *cognitive* capacity, our ability to perceive, to recall, to recognize, and to draw analogies, that is largely responsible for this pluralistic nature of our ontology. We shall encounter it in the *cognitive emergence* of new objects in the argument from cosmic evolution. That argument will be given first; it is in support of the notion of the unity of nature (substantive monism), an important notion which in a sense counterbalances the pluralistic notion of ontology.

2 THE ARGUMENT FOR SUBSTANTIVE MONISM FROM COSMIC EVOLUTION

The unity of nature can best be argued from cosmic evolution. In the process, the cognitive emergence of various branches of science will become evident.

Cosmology can tell us a lot about the early universe; sketching it in giant steps, the initial stages of the cosmic expansion first leads to the formation of elementary particles (in the first 10^{-35} seconds) followed by the formation of the lightest atomic nuclei (after the first minute); this is followed by the formation of the first atoms (after the first million years or so) and the appearance of very tenuous gas clouds. All this is studied as a branch of *physics*. So is the subsequent evolution of stars and galaxies (astrophysics). Also nucleosynthesis, starting with the reactions between light atomic nuclei in the interior of very hot stars and ending eventually in the production of the nuclei of all the chemical elements, belongs to astrophysics. But in regions where the temperature has become low enough, a new entity arises: molecules are beginning to form. These and their interactions are no longer described by physics; they are treated in a different branch of science, *chemistry*. Eventually, the solar system starts to evolve and the planet earth solidifies. There is a differentiation of new features of land and oceans, there is rock formation and vulcanic activity. The study of these belongs to *geology*. And as conditions arise for the stability of more complex organic molecules, the first primitive forms of life begin to evolve. Another new notion, living organisms, is associated with yet a different branch of science, *biology*.

One can of course continue in this way and eventually reach the complexity of the world as we know it today. All the branches of science will be encountered on the way. The conclusions from cosmic evolution are clearly the following:

(1) There is a unity of nature (a substantive monism) in that the fundamental entities evolve and unfold into increasing complexity. This process is *continuous*[1] and there are no new and extraneous entities added during that evolution. All objects are in last analysis different forms and manifestations of the entities originally present. There is therefore no room for a Descartian dualism not to speak of a substantive pluralism.

(2) This continuous development (cosmic evolution) is punctuated by the apparent emergence of new objects (atoms, stars, the earth, organisms, etc.) which call for *new concepts* (and associated new words). These objects are identified by our cognitive apparatus in terms of certain characteristic properties (attributes) suggesting that at this stage *something qualitatively new* has evolved. It is new because it differs perceptively in essential ways from anything that there was at any of the earlier stages; there is a recognition of this fact which is sudden despite the realization that nothing discontinuous has happened. I shall call this *cognitive emergence*.

(3) The characterization of these new objects and of their relations to other objects (their environment) is done by means of an *idealization* in which their detailed structure has become unimportant; that structure is either simplified

[1] Since our considerations must include quantum mechanics, the term 'continuous' requires qualification to 'quasicontinuous'. The discontinuities in quantum mechanics do not prevent predictability but they restrict it to a probabilistic one. Instead of predicting a unique final state, one predicts several possible states each with an exactly computible probability. This qualification will always be implied when we use the terms 'continuous evolution' and 'predictability'.

or ignored completely. Characterizations are therefore *approximations* in the sense that they have limited validity: beyond a certain observational precision they become empirically inadequate. This notion will be made more precise below.

(4) The simultaneous presence of different cognitive levels makes us lose sight of the common origin of the subject matter: the continuity of the evolution of nature becomes hidden when we are forced to look at its tremendous diversity *at a fixed time*. That fixed time view also hides the unity of science. What we call chemistry, biology, etc. are conventions due to particular *cognitive levels* associated with the fixed time view. Such conventions should not mask the substantive unity of nature (and therefore of science).

(5) Objects newly emerged during cosmic evolution themselves become the building blocks of yet other new objects. Again, idealization is used in their characterization (see also Shimony [1987]). We progress from fundamental particles to atoms, to molecules, to cells, to microorganisms, etc., each time ignoring certain structural details. A level structure of increasing 'coarseness' emerges and with it arise *levels of scientific enquiry*.[2] On each level we have an approximate description that is 'coarser' than the next level which describes component parts and which is therefore a 'finer' level. I prefer the terms 'coarser' and 'finer' level of theory because it does not have the connotations associated with terms such as 'more fundamental', 'superseding', 'supervening', 'primary', etc. which prejudice the case. The notions of finer and coarser will be made precise later on. However, this terminology is reasonable only in the physical sciences (our present subject matter) where size plays a leading role; for other scientific levels qualitative distinctions may dominate over quantitative ones. Consider for example the not unconnected 'levels': the geography, the economics, and the sociology of a given country.

3 WHAT IS MEANT BY A SCIENTIFIC THEORY?

Unfortunately, there are many discussions of theory reduction[3] without a clear statement of what is meant by a scientific theory. Since, furthermore, what I mean by 'scientific theory' differs from its usage in most accounts, it is necessary that I be quite explicit here.

Philosophers often refer to the history of science and accept as scientific theory anything that was so labelled historically. This usage does not distinguish between *stages of development* of a theory and the final theory. It permits such antirealist arguments as those found in Laudan [1981]. It also washes out the distinction between the trials and errors that are part and parcel of the long time labor by many hands which characterizes the development of a theory, and scientific revolutions which are relatively sudden steps to a different cognitive level. That lack of distinction resulted in some of the conclusions deduced by Kuhn [1970]. To give just one example, there is a long list of 'theories on electricity and magnetism' (Coulomb's, Ampere's, Faraday's, etc.) which are all stages in what eventually became the theory of (classical) electromagnetism; even Maxwell's 'theory' contained erroneous features (vortex 'theory' and ether). But it was a scientific revolution that brought about the development of quantum electrodynamics, a 'finer' theory that supersedes the 'coarser' classical electrodynamics.

---

[2] When we say that each theory refers to a specific level, we must exclude 'mechanics' in the general sense from the class of theories. 'Mechanics' is a frame that can serve many different levels depending on the objects and forces that are substituted. This holds for Newtonian mechanics as well as for quantum mechanics. Similarly, statistical mechanics interpolates between the levels of the micro-world and the macro-world by permitting us to treat very large collectives of the micro-world. An argument based on lattice structure which shows that quantum mechanics is not restricted to one level was given by Primas [1977] and [1981].

[3] We shall always use the philosopher's rather than the scientist's sense of 'reduce': S reduces to T if S can be derived from T. See Nickles [1973].

Philosophical questions on intertheory relations and specifically on theory reduction are properly addressed to the mature theories rather than to the successive developmental stages of a theory; the latter belong to the context of discovery. Eventually, only the mature theory needs justification.

I shall characterize 'scientific theory' in the following way. First, I distinguish sharply between empirical laws, conjectures, and theories. Empirical laws are inductively generalized observation statements usually cast into a mathematically simple form: Kepler discovered that the planetary orbits as inductively obtained from observations can be very well represented by ellipses in which one focal point is the location of the sun. Kepler's laws are not theories and neither are Galileo's laws of free fall and projectile motion.

While empirical laws provide esential justification for theories, they do not enter the question of reduction of one theory to another. They are 'reducible' to theories only in the sense that their mathematical representation (equations) may be derivable from theories: theories explain laws. Empirical laws and observations statements have *factual* ontological content (the existence of the sun and planets in the case of Kepler's laws) while theories have *formal* ontological content.

Conjectures are conceptual schemes with poor empirical support and are often without a mathematical framework; but they are often called 'theories'. The effluvial 'theory' of static electricity, phlogiston 'theory', 'theories' of spontaneous generation, and many others provide examples. Despite their names, they were still only conjectures.

Secondly, I distinguish between three kinds of theories although there is obviously no sharp separation between them: accepted, mature, and established theories. An *accepted theory* may be used by the majority of scientists because (at the time) it is the most plausible of the theories currently available on that subject; it is not necessarily accepted because scientists are convinced of its correctness. When an accepted theory is backed by a mathematical framework, has received sufficient empirical support, and coheres logically well with other theories (Rohrlich and Hardin [1983]), it may eventually evolve into a *mature theory*. That happens when it is believed to be (approximately) true by the vast majority of scientists on the grounds of very strong evidence. The qualification 'approximately' refers to the *validity limits* (see below). These limits are generally still unknown for a mature theory. Newton's gravitation theory before general relativity is a good example. But when the validity limits do become known (typically as a consequence of a scientific revolution), a mature theory becomes an *established theory* (Rohrlich and Hardin [1983]). In that case it is not discarded but continues its autonomous life within those limits.

Whether there have ever been *mature* theories which had to be discarded as a consequence of a scientific revolution is debatable; it depends on the line one draws between accepted and mature theories.

A common source of confusion is the presence of gratuitous conceptual appendages which occur in accepted or even in mature theories. Such gratuitous appendages are not required by the mathematical framework of the theory but are typical of the early stages of development of a theory. Examples include: absolute time and space (in Newton's mechanics), the ether, and most recently the left-right symmetry of all laws of nature (false for weak interactions). Eventually, such appendages are dropped without affecting the central theory when enough evidence against them has accumulated. Again, the 'theories' before and after dropping of these appendages are only stages of development.

The importance of a mathematical framework or structure in physical theories cannot be overemphasized. It provides for the derivation of the central equations and its *quantitative* explanatory and predictive power.

Perhaps even more important is the ability of the mathematical framework to probe where human intuition fails: when the theory refers to those aspects of nature which lie outside our direct experience, the mathematical structure M becomes the backbone of the scenario (the model) which characterizes this indirect knowledge. This ability of M gains increasing importance as investigations lead farther from everyday experience: the atomic and subatomic domains, the distant past in cosmology, distant galaxies, etc.

The conceptual model associated with a theory is largely derived by confronting M with empirical evidence and with neighboring theories (testing the coherence). It involves informal language and is not the result of logical-mathematical deduction.

Finally, the validity domain, D, of a theory must be made precise. The 'boundaries' of D are characterized by the relative magnitudes of physical variables. Typically, there exists a *characteristic parameter*, p, which is the ratio of two physical variables of the same dimensions of length, time and mass. The boundary of D is reached when p is no longer negligible. It is therefore not sharp but is given as a *known* error estimate (an order of magnitude estimate): it is the error one makes by applying that theory (rather than the finer one). The classic example is D(Newtonian mechanics) which is given by $p=(v/c)^2$. Here v is a typical velocity and c is the speed of light. The finer theory is special relativity.

It is evident from the above that we shall deal here only with mature or established theories. The characteristic components of one of those which are important for theory reduction are: its mathematical-logical structure, M; its domain of validity, D; its language and conceptual content, L (its central terms, its formal and informal language), its epistemic component, E; and its ontic component, O. A more complete list of theory components can be found in Bunge [1985].

4 MATHEMATICAL INTERLUDE

One of the central notions in the following arguments for a pluralistic epistemology and ontology is the emergence of new qualities when suitable limits are taken. To introduce this point, a mathematical analogy will be very helpful. The real numbers consist of two qualitatively different sets: the rational numbers, R, and the irrational ones, I. The latter cannot be written as the ratio of two integers, the former can. But for every number in I there exists an infinite sequence of numbers in R that converges to it. For example, there exists a sequence of elements of R,

$$3,\ 31/10,\ 314/100,\ 3141/1000,\ 31415/10000, \text{etc.}$$

Which converges in the limit to the irrational (in fact transcendental) number π. Furthermore, every element of I can be written as an infinite sum of elements of R. For example, $\pi=3{\cdot}14\ldots$ or

$$\pi=3+1/10+4/100+1/1000+5/10000+\ \ldots$$

All the numbers in that sequence are elements of R but in the limit there 'emerges' something *qualitatively* different: an element of I. Thus, one concludes:

(a) A limiting process can lead from one kind of numbers to a qualitatively different number.
(b) It is possible to approximate a number by a qualitatively different number (*e.g.* an irrational number by a rational one).
(c) An irrational number cannot be written as a function of rational numbers (nor *v.v.*), *i.e.* there is no finite set of algebraic manipulations that will do it; there is no way of mapping one set into or onto the other.

5 VIEWS ON REDUCTION

For a long time the strong reductionist view of logical positivism dominated. If theory S is to be reducible to theory T (*i.e.* when S can be completely deduced from T), then the central terms of S must be functions of the central terms of T. Such 'bridge laws' are necessary for a complete logical-mathematical deduction of S from T (see for example Schaffner [1967] and Suppe [1977]). Physicists, however, typically deduce only the *mathematical structure* of S from that of T and pay little attention to whether the concepts resulting from the physical interpretations of the symbols involved in those structures permit such a functional relation. They work largely intuitively. The mathematical structure or framework of the theory is considered to be primary, and the central terms (the meaning of certain central symbols) are later derived from the applications of that framework to actual situations. This weaker view on reduction allows for meaning change of terms (Feyerabend [1962] and [1970]) without affecting this reductive relationship (see Section 7 below).

At the opposite extreme to complete reduction are those who opt for concept *replacement*. This view holds that the coarser (superseded) theory is simply wrong and is to be replaced in its entirety by the finer one. But that goes too far as will be argued below. In any case, it can be stated with little doubt that few scientists would accept the replacement of electromagnetic field concepts by those of mechanics as suggested by Spector [1978]. Some arguments against the replacement view (with respect to the ontology of a theory) have already been presented in Section 1.

Logical incompatibility between the central terms would indeed lead to contradictions between S and T and to the impossibility of even weak theory reduction. This, in turn, would be a strong argument for the lack of cumulativity of knowledge. But I shall argue that there can be no logical incompatibility (contradiction) because S and T refer to different cognitive levels. The mathematical framework of S is rigorously derived from that of T (a derivation which involves limiting procedures); but the interpretations and the ensuing ontologies of S and T are in general not so related. They involve qualitative differences and for this reason demand independent recognition. In this way, one comes close to Feyerabend's theoretical puralism and at the same time one ensures a well-defined logical-mathematical linkage between S and T.

6 REDUCTION OF THE MATHEMATICAL COMPONENT

Reduction of a coarser theory S to a finer theory T requires that the mathematical framework M(T) implies M(S) when the validity domain D(T) is restricted to D(S). This restriction involves a *limiting process* which has much in common with the mathematical example just presented. Let M(coarse) and

M(fine) correspond to the rational and irrational numbers, respectively. Then M(coarse) approximates M(fine) like 3·14 approximates π. The error made can be estimated in order of magnitude. Furthermore, there are qualitative differences between S and T just as there are qualitative differences between rational and irrational numbers.

Finally, (no longer in analogy with the mathematical example above) the limiting process relating the Ms of the two theories is carried out by taking the limit to zero of the characteristic parameter p. M(fine) will be a function of p but M(coarse) will not depend on p. The reduction of the Ms can be written as

$$(p\rightarrow 0)\ \lim M(\text{fine}) = M(\text{coarse}).$$

Because M(coarse) no longer depends on p, some physical quantity present in the fine theory is absent in the coarser theory.

Given the finer theory, it is not obvious what the characteristic parameter p actually is. It becomes evident only when the coarser theory is known. This knowledge is of course present in the usual historical development of physical theory. However, there are exceptions. This happens in the *chronological inversion* of theory development, when the finer theory is known *before* the coarser one: in astrophysics we know the fine theories (atomic physics, chemistry) while the coarser theories (formation of stars and galaxies for example) are in a rather rudimentary stage. And there is no way to deduce these theories from the 'finer' theories since the parameters p are still unknown. We have here another restriction to the claim for superiority of the finer theories.

The observation that one cannot construct an unknown coarser theory from the finer one even though knowledge of the coarser one permits its reduction to the finer one is not new (Anderson [1972]). But this observation now fits into a larger scheme.

Ignoring the limiting process has resulted in a misunderstanding of the process of reduction. The coarser theory cannot be written as a *function* of the finer one even in its mathematical formulation. This is most apparent in the mathematical analogue of Section 4, conclusion (c): rational numbers are not functions of irrational ones.

If the finer theory is endowed with certain symmetry properties then this limit will produce the coarser theory with *its* symmetry properties. The reduction of the symmetry properties of M(coarse) to those of M(fine) plays a very significant role: it determines the change in the invariance properties of M. From a technical point of view, it shows that the limiting process is highly nontrivial and must be carried out very carefully: the symmetry reduction may be the result of group contraction, and the limit can only be carried out in suitable group representations. But we shall not pursue these mathematical matters here.

As a well-known example consider the characteristic parameter $p = (v/c)^2$

for the validity domain of Newtonian mechanics, NM. It is reduced to special relativistic mechanics, SR, by

$$((v/c)^2 \to 0) \lim M(SR) = M(NM).$$

In this limit, the speed of light c disappears from the theory and the spacetime symmetry of SR, Lorentz invariance, becomes the spacetime symmetry of NM, Galilean invariance.

Another example is the reduction of geometrical optics, M(GO), to electromagnetic theory, M(EM). The characteristic parameter p is here the ratio of the wave length of light to the typical size of the objects considered. In the limit, as p goes to zero, the fundamental equations of GO are derived from those of EM. The ray optics of GO is a limit of the wave optics of EM; at the same time, the explicit dependence on the wave length disappears from the equations of the theory.

Finally, Bohr's correspondence principle is nothing but a reduction postulate. It requires the reduction of classical mechanics, CM, to quantum mechanics, QM. As far as their Ms are concerned, they are related by a limiting process $p \to 0 \lim M(QM) = M(CM)$ where p is (roughly speaking) the ratio of the size of a single quantum of some observable in QM to its size in CM. It is dimensionless but proportional to Planck's constant h. In the limit, the theory becomes independent of Planck's constant h.

These examples of course do not prove that such a reduction is always possible. The point is, however, that if a mature coarse theory S is superseded by a finer theory T in such a way that S survives within a certain domain D (and thus becomes an established theory) then it is a *necessary* requirement for vertical logical coherence of theories (see Rohrlich and Hardin [1983]), that such a reduction exists. This means that there must exist a parameter p which permits the above limit to be carried out.

The mathematical details of such reductions have been worked out on various levels of rigor. A recent rigorous study of the reduction of Newton's to Einstein's gravitation theory is given by Kunzle [1976] and Ehlers [1986].

It is also possible that a theory involves more than one parameter so that their limits lead to several different theories. For example, relativistic quantum electrodynamics has one limit to nonrelativistic quantum electrodynamics and another to relativistic classical electrodynamics (see Rohrlich [1965] chapter 8 in this context).

The domain D of the coarser theory is clearly fully contained within that of the finer theory. It is given by the characteristic parameter which provides the error estimate; if that error is negligible, one is within D(coarse).

7 THE LANGUAGE COMPONENT

There are three things that may happen to the central terms of the finer theory in the limit: their meaning may remain unchanged, they may undergo a meaning change, or they may disappear altogether. On the other hand, new central terms may arise in the coarser theory that did not exist in the finer theory. I do not know of a single nontrivial case of theory reduction where *all* central terms remain unchanged (homogeneous reduction according to Nagel [1961]).

How does the meaning of central terms come about? The mathematical framework, M, of a newly proposed finer theory typically involves symbols which may at least tentatively have the same meaning as in the coarser theory; it also involves new symbols whose meaning is yet to be clarified. These symbols belong to the formal part of the language L of the theory. As the new theory is being applied to specific cases, one arrives at an interpretation of the various symbols: they are given a meaning. This process necessarily involves the informal part, the semantics, of L. Eventually, one develops a conceptual picture, a model. Such a model may sometimes take years to complete: sixty-odd years after the development of quantum mechanics, there still is no complete agreement on the meaning of ψ.

Of course, there is no reason to expect this semantics of the new theory to be the same as that of the old one. The difference is part of what is usually dubbed 'deeper insight'. The new theory is said to provide finer details because it often tells about the structure or composition of the objects in the older (coarser) theory). In this process of developing a semantics for the finer theory, a great deal of the qualitative features which are present in the semantics of the coarser theory, disappear, and a new and different cognitive level is being established.

Conversely, comparing the coarser theory, S, with the finer one, the difference in the semantics can be blamed on the *idealization* brought about by the limiting process that leads from M(fine) to M(coarse) in which certain features are neglected becucase of the assumption $p=0$. It is therefore not surprising that the interpretation of S provides less detail or structure; an information loss seems to have occurred: the coarser theory does not seem to 'know as much' about the subject matter as the finer one. A more accurate description, however, would be to say that it 'knows different things'. The explanation offered by T is on a deeper level and in that sense better than the one offered by S (within its domain). But the two theories may use quite different central terms and those of S may be easier to understand because they are closer to everyday concepts compared to those of T. It is perhaps for this reason that Kuhn [1970] speaks of a loss of capabilities (presumably meaning explanatory power) of the finer theory relative to the coarser one. The terms are on different cognitive levels and for this reason necessarily convey conceptually different explanations. There is no loss logically (formally) but there is one informally. For example, the intuitive picture of 'heat flow' is lost on the microscopic level provided by statistical mechanics.

The main goal of language is a pragmatic one: to communicate; that's where ordinary and informal language enters. The goal of logical reasoning on the other hand requires only formal language; but that's not enough for interpretation, recognition, and description. What is taken as 'primitive' varies from one cognitive level of description to another. And for a 'complete' description all levels are absolutely essential: a description of the furniture of this room in terms of *atoms only* is cognitively incomplete!

Consider these ideas in the context of physical theories. As an example of the qualitative difference of central terms of two theories S and T, consider time in Newtonian mechanics, t(NM), and time in special relativity, t(SR). They differ in essential ways, the former being common to all ('public'), the latter being different for different reference frames ('private'). Nevertheless, as the parameter p becomes smaller and smaller, t(NM) becomes a better and better approximation to t(SR) because the differences between the private times become smaller and smaller. In the limit, t(NM) becomes 'exact'. A central term approximates a qualitatively different one.

The 'logical incompatibility' proposed by some (Feyerabend [1962] and [1970]) and rightly objected to by others (Yoshida [1973]) is only a cognitive incompatibility. The *logical coherence* between the finer and the coarser theories is ensured by the reducibility of M(S). The limit involved there ensures logical *continuity* and compatibility.

In analogy with the mathematical example in Section 4, conclusion (b), a similar statement can be made about the central terms of theories: some of the central terms of S result from limits of descriptions in T and are therefore capable of approximating them successfully.

Examples: (1) A mirror is idealized as 'perfectly smooth' when we know very well that the molecular structure of glass does not allow for such a concept and requires at least a granular surface. The size of the molecules is taken here to the limit zero. (2) The central term '*light ray*' in geometrical optics, GO, is seen to emerge from a fine structure of a very complicated superposition of *waves* in electromagnetic theory, EM; in the limit of vanishing wave length, this superposition becomes a ray. A well-known related phenomenon is the sharp shadow in GO which has as its fine structure in EM the phenomenon of diffraction on a sharp edge. The sharp shadow and the diffraction do not contradict one another; the shadow is sharp in the limit.

These examples show that despite the *continuous* change from one M to the other in the p→0 limit, the interpretation is *discontinuous* in the limit. This discontinuity is due to the *cognitive emergence* of a qualitatively new and different description. The equations obtained in the limit have a new and different interpretation (semantics).

Finally, in analogy with conclusion (c) of Section 4, one has the impossibility of expressing every central term of S as a function of the terms of T. A well-known analogy from visual perception may illustrate this point better than physical examples. Let the finer 'theory' be a distribution of points on a sheet of paper (a dot matrix). The density of these points varies in a complicated way. When viewed from some distance, however, the individual points are no longer discernible. The density of points becomes a continuous distribution of different shades of gray. As the distance from the paper is gradually increased, these shades conspire to make us suddenly(!) recognize the face of Winston Churchill. The notion of that face was of course foreign to the finer description but was previously known to us. The terms of one description gave way to terms of a qualitatively different one; yet, there was a continuous change in the observation process (distance from the viewer). And it is obvious that the finer description (the dot matrix) cannot convey the concept of the coarser one.

In a similar way, the wave picture of light gradually changes as one approaches the geometrical optics limit until a cognitive act discontinuously recognizes a ray picture. And again, the wave picture cannot express what the ray picture tells us.

Cognitive emergence is here based on previously known features as one goes from the finer to the previously known coarser theory. On the other hand, in the construction of finer theories (in scientific revolutions) new features are conjectured by the imagination and originality of the scientist. There are thus two different kinds of 'cognitive emergence'.

Also, a sharp distinction must be made between meaning change of a term associated with vertically adjacent scientific theories (reductive change) and meaning change of a term in a nontechnical or pre-scientific theory to that in a mature scientific theory (historical change): 'heat' underwent considerable historical change from caloric theory to thermodynamics. Historical meaning change must be kept clearly separate from reductive meaning change.

One concludes that a *logical* deduction of L(coarse) from L(fine) is not possible! Whether one wishes to deny reducibility for this reason or accept the mathematical deduction of M as sufficient reason for calling a theory 'reducible', now becomes a matter of definition.

8 PLURALISTIC EPISTEMOLOGY

A scientific theory attempts a rational and explanatory description that is fully consistent with empirical results *within the approximation* provided by its validity limits (empirical adequacy). In addition and in contradistinction to the constructive empiricist view of van Fraassen [1980], the resultant *approximate truth* receives additional strong support from the horizontal and vertical logical coherence with other theories (Rohrlich and Hardin [1983]). Logical coherence adds very substantially to the rational belief in its (approximate) truth and provides for an extreme 'robustness' of the theory (Wimsatt [1981]).

This kind of approximate truth of a theory must not be confused with the approximate description of an actual physical phenomenon which is in general extremely complex. Ronald Laymon [1982] has elaborated on this point.

Given that each theory T has its own E(T) and D(T), one is likely to commit the common fallacy of assuming that the finer theory is always more true than the coarser theory. The earlier example of the dot matrix being unable to describe (on *its* level) the face of Churchill exposes this fallacy. The finer theory fails completely in accounting for qualitative features easily described by the coarser theory. In this case, there is an epistemic loss when one restricts oneself to the *finer* theory. Truth is here established by correspondence to different cognitive levels, each making its own contribution. It follows that *the qualitative character of the coarser theory demands recognition in its own right* despite the knowledge of the finer theory. Since only the mathematical framework can be derived in the reduction process and not the interpretation or the epistemic component, there is consequently a demand for a pluralistic view of the epistemology of physical theory. The key function of the finer theory is to provide the validity limits for the coarser theory and thus for its truth. But since the finer theory also has its validity limits, its truth is also only an *approximate truth*. No ultimate theory being known, there is no ultimate (exact) scientific truth known.

Physical examples: (1) The geometric shape of crystals (not their symmetry) are not found on the level of atomic physics but emerge holistically due to idealizations (smooth surfaces, sharp edges, etc.). (2) Theories of collective or cooperative phenomena in any field of natural science provide examples of coarser theories not being replaceable by finer theories which deal with the individual objects (Shimony [1987]). This kind of holism is of course again the result of a change of the universe of discourse, a change from one cognitive level to another which is qualitatively different.

Our knowledge is true knowledge within the appropriate validity domain and on a particular cognitive level of idealization. But it necessarily involves only approximate truths. There is nothing inconsistent about that. In fact, these truths *complement* one another, each theory making a contribution on its own level. The scientific revolution brought about by the development of a finer theory *adds* to our knowledge; it does not make the surviving coarser theory unnecessary or useless; nor does it call for the replacement of the old theory within its domain; it makes science cumulative. In this way, one is led to a *pluralistic epistemology of science*.

As an immediate consequence, it is no longer compulsory to ask for 'convergence to an ultimate exactly true theory'. Rather, there is a convergence to a *logically coherent net of theories* each with its own M and associated interpretation (see Tisza [1963] for an attempt to construct such a net in a different context). The connectedness of this net expresses the logical coherence of theories. Pluralistic epistemology does not imply a disjointness of nature or science; on the contrary, it affirms the unity of nature (our substantive monism based on the argument from cosmic evolution in Section 2) by the logical coherence of the theories of the net, and thereby it also confirms the unity of science. But there are different cognitive views within this net, each in its own right.

9 PLURALISTIC ONTOLOGY

The formal part of a scientific theory does not contain the empirical information which supports it. It can therefore only make formal ontological claims. The empirical support, in particular the laws obtained from observation or experimentation, involve factual ontological claims, and they must be 'matched' with the formal claims of the theory.[4] After they are matched we speak simply of the ontology of the theory. The matching provides the justification for asserting the existence of physical objects and their predicates as predicted by M and L. This necessarily requires going beyond the formal language and using the informal component of L.

The external reality thus ascertained involves observables which can be 'seen' by us (no matter how indirectly!) *on the specific cognitive level for which the theory has been constructed.* The degree of indirectness can be quite large: the technique of tunneling microscopy is highly indirect, but one *can* observe individual atoms that way. Observables are thus independent of the technology available for observation, independent of the kinds of instruments used, and independent of the mode by which the observation is performed.

Is there a reduction of the ontological component O(coarse) from O(fine)? Such a reduction requires a derivation of O(coarse) from O(fine). I claim that this is in general not possible. Nor can one deduce O(coarse) from M(fine).

Consider first an example: Einstein's gravitation theory. It claims the existence of curved space but not of gravitational forces. In the Newtonian limit, however, the interpretation of the mathematical formalism leads to a flat space and to gravitational forces. Which ontology is correct?

The two ontologies are associated with different cognitive levels. On each level the theory gives us a different 'model of the world' which has *characteristic features that cannot be expressed by any other level including the immediately superseding one*. The models thus complement one another.

Why can the ontology of one theory O(S) in general not be expressed in terms of the ontology O(T) of the superseding one? A formal deduction of M(S) from M(T) is assumed to be possible. But O(T) is obtained from the interpretation of M(T) using informal language since it is not a *logical* deduction; similarly, O(S) is obtained from M(S) and therefore also from M(T) not by logical deduction. There is therefore no logical (formal) deductive path for O(S). And a direct 'mapping' of O(S) into O(T) is barred by the cognitive differences of the two levels: there is no way to express the macroscopic properties of a table in terms of atoms and their properties.

---

[4] This does not mean that the observation statements are entirely free of theoretical notions. But their 'theory-ladenness' is sufficiently weak not to invalidate the substantial support a theory can receive from empirical results.

What there is can be known to us by means of our cognitive apparatus on different levels, and so can its attributes and its lawful behavior. These cognitive levels are qualitatively different and thus complement one another even though they are hierarchically related in a logical way.

The permanence of established theories, the nature of their truths (approximate within limits that are accessible to quantitative estimates), and the limited reducibility of theories (the mathematical formalism only) leads not only to cumulativity of scientific knowledge but also to a pluralistic epistemology and ontology.

Syracuse University, Syracuse, New York

REFERENCES

Anderson, P. W. [1977]: 'More is different', *Science*, 177, pp. 393–6.

Bunge, M. [1985]: *Treatise on Basic Philosophy*, vol. 7 (Part I). Dordrecht: Reidel.

Cartwright, N. [1983]: *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Ehlers, J. [1986]: 'On Limit Relations Between, and Approximative Explanations of, Physical Theories', in B. Marcus *et al.*, *Logic, Methodology and Philosophy*, pp. 387–403. New York: Elsevier Science Publishers.

Feyerabend, P. K. [1962]: 'Explanation, Reduction and Empiricism', in H. Feigl and G. Maxwell, *Minnesota Studies in the Philosophy of Science*, vol III, pp. 28—97. Minneapolis: University of Minnesota Press.

Feyerabend, P. K. [1970]: 'Problems of Empiricism', in R. G. Colodny, *The Nature and Function of Scientific Theories*, pp. 275–353. Pittsburgh: University of Pittsburgh Press.

Kuhn, T. S. [1970]: *The Structure of Scientific Revolutions*, 2nd ed. Chicago: Chicago University Press.

Kuenzle, H. P. [1976]: 'Covariant Newtonian Limit of Lorentz Spacetime', *General Relativity and Gravitation*, 7, pp. 445–57.

Laudan, L. [1981]: 'A Confutation of Convergent Realism', *Philosophy of Science*, **48**, pp. 19–49.

Laymon, R. [1982]: 'Scientific Realism and the Hierarchical Counterfactual Path from Data to Theory', *PSA 1982*, vol. 1, P. D. Asquith and T. Nickles (*eds*.). Philosophy of Science Association, East Lansing, pp. 107–21.

Nagel, E. [1961]: *The Structure of Science*. New York: Harcourt, Brace & World.

Nickles, T. [1973]: 'Two Concepts of Intertheoretic Deduction', *Journal of Philosophy*, 70, pp. 181–201.

Primas, H. [1977]: 'Theory Reduction and Non-Boolean Theories', *Journal of Mathematical Biology*, 4, pp. 281–301.

Primas, H. [1981]: *Chemistry, Quantum Mechanics and Reductionism*. Berlin: Springer-Verlag.

Rohrlich, F. [1965]: *Classical Charged Particles*. Reading: Addison-Wesley.

ROHRLICH, F. and HARDIN, C. L. [1983]: 'Established Theories', *Philosophy of Science*, 50, pp. 603–17.

SCHAFFNER, K. F. [1967]: 'Approaches to Reduction', *Philosophy of Science*, 34, pp. 137–47.

SHIMONY, A. [1987]: 'Methodology of Synthesis: Parts and Wholes in Low Energy Physics', in Robert Kargon and Peter Ackinstein (eds.), *Kelvin's Baltimore Lectures and Modern Theoretical Physics*, pp 399–423. Cambridge: M.I.T. Press.

SPECTOR, M.' [1978]: *Concepts of Reduction in Physical Science*. Philadelphia: Temple University Press.

SUPPE, F. [1977]: *The Structure of Scientific Theories*, 2nd ed. Urbana: University of Illinois Press.

TISZA, L. [1963]: 'The Conceptual Structure of Physics', *Reviews of Modern Physics*, 35, p. 151.

VAN FRAASSEN, B. C. [1980]: *The Scientific Image*. Oxford: Clarendon Press.

WIMSATT, W. C. [1981]: 'Robustness, Reliability, and Overdetermination', in M. B. Brewer *et al.*, *Scientific Inquiry and the Social Sciences*, pp. 124–63. San Francisco: Jossey-Bass Publishers.

YOSHIDA, R. M. [1977]: *Reduction in the Physical Sciences*. Halifax: Dalhousie University Press.

Brit. J. Phil Sci. 39 (1988), 313–329 *Printed in Great Britain*

*Metaphysics, MSRP and Economics**

J. C. GLASS AND W. JOHNSON

ABSTRACT

Lakatos' MSRP is utilized to provide a response to Koertge's claim (in her 'Does Social Science Really Need Metaphysics?') that the heuristic significance of metaphysics has been vastly overrated. By outlining the hard cores and positive heuristics of the two major research programmes in economics (namely, the 'orthodox' and 'Marxist' research programmes), the paper demonstrates (in opposition to Koertge's claim) not only that the metaphysical statements in the respective hard cores are far from vague but also how these exert an important regulative influence both on the theories and economic policy recommendations generated within the respective programmes. It also indicates how the adoption of the MSRP method of appraisal in economics would help to protect against the danger of entrenched metaphysical dogmatism with its implications not only for economic policy recommendations but also for moral and political recommendations.

I INTRODUCTION

Does social science really need metaphysics? This interesting question has been discussed and answered by Koertge [1979a]. Her answer is:

> *Yes, if used with discretion* ... 'yes' because metaphysical questions are so very interesting and there is no doubt that important scientific questions are raised by philosophical debates ... and the competition between metaphysical research programmes ... 'if used with discretion' because the possibilities for the *rational* influence of metaphysics on the content of a scientific theory are rather limited and because there is no efficient internal method for criticizing one's metaphysical framework. ([1979a], p. 32, her italics)

More specifically, Koertge has argued for discretion in the use of metaphysics in social science because (i) the typical vagueness of metaphysical statements severely limits the extent to which metaphysics can exert a useful regulative influence on the construction of new theories by ruling out a certain range of theoretical possibilities ([1979a], pp. 22–8); and (ii) the difficulty of internally criticizing a metaphysical framework means there is a very real danger that a

---

* We appreciate the assistance given by Professor N. Koertge, who provided us with translations of her articles, and the helpful comments of an anonymous referee.

situation of entrenched metaphysical dogmatism will arise whenever a particular metaphysical framework becomes a fad ([1979a], pp. 28–33).

In this paper we utilize Lakatos' methodology of scientific research programmes (or MSRP, for short) to provide a response (based upon our experience of economics) to certain aspects of Koertge's discussion and conclusion. In particular, our response (which draws upon Glass and Johnson [1988]) argues that (1) the metaphysical statements contained in the major research programmes in economics are by no means vague; and (2) these metaphysical statements do exert a considerable regulative influence on theory construction within the respective economics research programmes. Given (1) and (2), we further argue that the adoption of the MSRP method of appraisal in economics would serve the useful purpose of (3) stimulating an increased awareness of the importance of metaphysical competition and hence encouraging the metaphysical pluralism that Koertge advocates as 'the only hope for avoiding metaphysical dogmatism' ([1979a], p. 30); and (4) making explicit the important link between on the one hand the specific metaphysics associated with a given economics research programme and, on the other hand, the theoretical predictions and economic policy recommendations generated within that research programme.

2 METAPHYSICS AS SPECULATIVE, UNTESTABLE WORLD PICTURES

Because of the difficulties in providing a precise characterization of metaphysical statements (see Koertge [1979b], p. 242 on Watkins, [1975]), Koertge views metaphysical statements as statements which provide speculative, untestable world pictures. Hence, for Koertge, metaphysical statements have two notable characteristics:

(1) 'they are statements which make claims about the fundamental nature of the domain under investigation' and
(2) 'there is no way of testing them directly against experience' ([1979a], pp. 2–3).

Given this view of metaphysics, the unfalsifiable, distinctive hard core of a Lakatosian research programme can then be viewed as providing an unfalsifiable, conjectural world picture. Hence in MSRP we have metaphysics as hard core (see Koertge [1979a], p. 20). We shall utilize this view of metaphysics as hard core in the subsequent discussion of research programmes in economics.

3 THE ROLE OF METAPHYSICS IN THE DEVELOPMENT OF NEW THEORIES

Koertge agrees with Agassi [1964] 'that metaphysics is indeed cognitively relevant *to* scientific inquiry' but disagrees with Lakatos that metaphysics should 'play a central explanatory role *within* science' (see Koertge [1979a], p. 22, her italics). Thus while Koertge agrees with Agassi that metaphysics is an important source of scientific problems, in the sense that many 'scientific theories are deliberately designed to be tentative, partial answers to metaphysical questions' ([1979a], p. 13), she argues (in disagreement with Lakatos) that metaphysics should be given a much less significant role within scientific inquiry.

One of Koertge's reasons for playing down the role of metaphysics within social science is her view that the *regulative influence* of metaphysics in the development of new theories is much less important than is generally claimed. In opposition to Koertge's view that the regulative influence or 'heuristic significance of metaphysics has been vastly overrated' ([1979a], p. 22), it is our contention that, where the social science of economics is specifically concerned, metaphysics *has* a most important regulative influence on the development of new theories. In order to support this contention, we shall first outline what we consider to be the two major research programmes in economics. This outline will then be utilized to demonstrate not only that the metaphysical statements contained in these two major economics research programmes are by no means vague, but also how these metaphysical statements exert a considerable regulative influence on theory construction within the respective economics research programmes.

(Note, in providing this demonstration, we are not hypothesizing that the regulative influence of metaphysics on theory construction is greater in economics than in, say, the 'natural sciences'. Rather, since Koertge has made a specific claim with regard to social science, our demonstration relates solely to the social science of economics.)

4 MAJOR RESEARCH PROGRAMMES IN ECONOMICS

When we start to look for major research programmes in economics, which attempt to explain the various economic aspects of advanced capitalist economies, two obvious candidates spring quickly to mind: the 'orthodox' economics research programme (containing the numerous microeconomic and macroeconomic theories which are predominantly taught in western universities, polytechnics and schools) and the 'Marxist' economics research programme (containing both Marx's economics and subsequent developments of Marx's economic analysis). These two research programmes not only have a long history but also have distinctive hard cores and positive heuristics. In addition, various subprogrammes may be identified within each of these research programmes, though that need not concern us here. (For support that the orthodox and Marxist economics research programmes can be viewed as the two major research programmes in economics, see Archibald [1979], p. 309.)

In the following discussion we shall first outline (in Section 4.1) what we consider to be the hard core of the orthodox economics research programme and then (in Section 4.2) the hard core of the Marxist economics research programme, with the latter outline emphasizing not only the distinctive differences between the respective hard cores but also the non-vagueness of the metaphysical statements constituting these hard cores. This discussion is followed (in Section 4.3) by a brief statement of the connection between the positive heuristic and the hard core of a research programe and (in Sections 4.4 and 4.5) by an outline of the positive heuristics of the orthodox and Marxist economics research programmes, with the discussion in Section 4.5 indicating how the distinctive differences in these positive heuristics result from the way in which each positive heuristic is influenced by its own particular hard core. Section 5 will then provide a discussion of the regulative influence of metaphysics on theory construction in economics research programmes.

4.1 *The Hard Core of the Orthodox Economics Research Programme*

In order to appreciate how we view the orthodox research programme, note that the hard core of this programme, and hence the hard core underlying *both* (orthodox) microeconomic and (orthodox) macroeconomic research activity, is envisaged as being composed of the following four basic assumptions or hypotheses:

Individualism: It is assumed or hypothesized that economic analysis will rest on propositions about *individual* behaviour, with the individual economic agent being conceived in abstraction from his social and historical setting.

Rationality: It is assumed or hypothesized that the behaviour of each economic agent is *rational*. Essentially this assumption embodies the requirement that agents make their economic calculations in a consistent way. Hence, given a particular problem-situation, rational behaviour is viewed as that behaviour which is appropriate to this given situation in the sense that each economic agent acts in a way that is consistent with his given preferences. When taken in conjunction with the assumption of individualism, this means that the consistent behaviour of economic agents is exclusively related to the preferences of individuals, with the characterization of these preferences being independent of the individual economic agent's social and historical setting.[1]

---

[1] If an individual's preference ordering is represented by a utility function, then he chooses rationally when he chooses what he prefers, or when he chooses so as to maximize utility. While the notion of rationality given in the text is that which is commonly used in orthodox economic analysis, it should be noted that the concept of rationality has been the subject of considerable discussion (*e.g.*, see Hahn and Hollis [1979]). Also, note that Lawson [1985] argues that orthodox economic analysis has neglected Keynes' viewpoint in which the behaviour of individual economic agents and the social situation interact with each other so as to give rational behaviour both a social and a historical dimension.

Private property rights: Individual economic agents are assumed or hypothesized to have two kinds of *property rights*. The first kind relates both to the individual agent's mental and physical abilities and to his own labour time, while the second kind relates both to some amount of goods and services that either have been produced or will be produced and to some amount of natural resources.

Market economy: The rational behaviour of individual economic agents is assumed or hypothesized to take place in a *market economy*. In such an economy, economic agents (with their respective property rights in terms of labour time and in terms of goods, services and natural resources) are characterized as taking individual decisions with regard to the buying and selling of goods, services, natural resources and labour time.

The hypotheses listed under the above four headings represent the basic hypotheses underlying the orthodox research programme in economics. As such, this hard core of hypotheses forms the basis from which (orthodox) economic analysis is to develop.

(Note that while introductory economics textbooks may define macroeconomics as the study of the overall aggregates of the economy, in contrast to microeconomics which may be defined as the study of individual economic units within the economy, more advanced macroeconomic analysis indicates that macroeconomic theory, in common with microeconomic theory, rests on propositions about the rational behaviour of individuals. This, in turn, explains why individualism and rationality can be regarded as hard core hypotheses in relation to both microeconomic and macroeconomic analysis. Furthermore, since macroeconomic analysis also presupposes both private property rights and a market economy, we can see why the hypotheses of private property rights and a market economy can be regarded as hard core hypotheses in relation to both microeconomic and macroeconomic analysis.)

4.2 *The Hard Core of the Marxist Economics Research Programme*

The hard core of the Marxist research programme is very different from that of the orthodox research programme:

The social nature of individual economic agents: While Marx's analysis of capitalism deals with the behaviour of individual economic agents, he maintains that one cannot deal with the individual economic agent in abstraction from his social and historical setting. Consequently, Marx's analysis, in contrast to orthodox economic analysis, stresses the *social* nature of the individual economic agent within a specific historical setting. However, this does not mean that society is viewed as an entity distinct from the individual economic agents or that the individual economic agent is viewed as a product of society. Rather, Marx's social image of the individual economic agent means that there is an integration and interaction of individual and society in his economic analysis, so that change in individual economic agents is thereby also change in society and change in social circumstances is also change in individual economic agents (see Avineri [1968], pp. 86–95).

Rationality with a social dimension: The Marxist research programme, in common with the orthodox research programme, contains the basic assumption or hypothesis of *rationality*. However, the Marxist research programme views this assumption in a way that is distinctly different from that of the orthodox research programme. Thus, while this assumption embodies the requirement that economic agents make their economic calculations in a consistent way, this consistency of behaviour in the Marxist analysis is viewed as taking place in a historically specific social context. Hence, given a particular problem-situation, rational behaviour is taken as that behaviour which is appropriate to this given situation in the sense that individual economic agents act in a way that is consistent with their respective social relations (and therefore with their respective class positions as defined by these social relations), with these social relations corresponding, in turn, to the historically specific economic structure (in our case mature capitalism) under consideration.

Property ownership as a social relation: The Marxist research programme also views the basic assumption or hypothesis of *private property rights* in a way that is distinctly different from that of the orthodox research programme.

In Marx's analysis of capitalism the means of production are monopolized as the private property of the capitalist class. In contrast, the working class is presented as having no property rights over the means of production. This particular structure of property ownership, in turn, not only defines a specific social relationship between the two classes but also means that the production process has an important social character. On the one hand, it is the worker's non-ownership of the means of production which compels him to sell his only asset, his labour power, to the capitalist. On the other hand, it is the capitalist's ownership of the means of production which not only leads him to buy the worker's labour power but also empowers him to appropriate the surplus value created by the worker. In other words, the capitalist's ownership of property is not just simply viewed as an initial endowment of 'things' (such as buildings, machinery and raw materials) but rather as a *social relation* which, in turn, determines the social form of the production process.

Moreover, in Marx's analysis, since the structure of property ownership enables the capitalist not only to dominate the worker (via the former's ownership, and hence control, of the production process) but also to exploit the worker (by appropriating the surplus value), then the class antagonism which arises directly out of the capitalist production process can be crucially related to the specific structure of ownership and non-ownership of property that characterizes mature capitalism.

In contrast, the orthodox research programme generally assumes some initial structure of property ownership without attempting to analyse the specific social relationships which flow from these. Hence, in the orthodox research programme, property ownership is viewed essentially as ownership of 'things' (which may either be used in the production of further things or exchanged for other things) rather than as a social relation. Also, in contrast to Marx's analysis where the property rights have a *historical* character (since they are the property rights *specific* to mature capitalism), property rights in the orthodox research programme are *ahistorical* in character.

Social relationships between people lie beneath the market exchange of commodities: In the Marxist research programme, as in the orthodox research programme, a *market economy* is assumed. However, in Marx's analysis, the concept of exchange is very different from that of the orthodox research programme. Whereas in the latter exchange is simply a relationship between 'things' (with 'things' referring to the commodities which are exchanged via the market system, and where the commodities, in turn, refer not only to physical goods but also to labour time which workers sell as a good in the market), in Marx's analysis this relationship between things is simultaneously a relationship between people.

Thus, in Marx's analysis, the social character of production (that is, the social or class relationships which characterize the capitalist production process) is expressed through the exchange of commodities. Consequently, according to Marx, it is necessary to probe beneath the 'appearance' of mere exchange of commodities in the market to the 'reality' of social relationships between people. In his view, it is only by going beneath market exchange to the social form of production (and hence to the class relationships which characterize the production process) that exchange in a capitalist economy can be adequately analysed. Hence, in his analysis, the emphasis is heavily placed on the social production relations rather than on the phenomenon of market exchange. Moreover, since the social form of production lying beneath market exchange is taken to be that which is historically specific to mature capitalism, then exchange in Marx's analysis has a specific *historical* character as well as a *social* dimension.

Historical materialism: The growth of the forces of production (defined as the physical means of production, such as machinery, raw materials and buildings, plus labour power) is given a primary explanatory role in Marx's economic analysis. Thus, while at a particular stage in the historical development of an economy there is both a correspondence and a compatibility between the forces of production and the social relations of production (with the latter, in linking productive forces and economic agents in the production process, against the specific background of economic ownership and non-ownership of the means of production, defining the class positions of individual economic agents), the forces of production are given a dominant role. Hence, the forces of production are viewed as determining the social relations of production.

As the productive forces develop, they become incompatible with the existing social relations of production. Hence, rather than enhancing increased material production, the existing social relations now constitute an obstacle to the continued development of society's productive capacity. The growth of the productive forces is thus viewed as leading to a contradiction between the forces and relations of production and therefore to class conflict. The intensification of this incompatibility and conflict is then regarded as a driving force which leads to the breakdown of the existing mode of production (defined as encompassing the technical nature of the productive process, the social relations of production and the attendant property relations) and to its replacement by another mode of production (with new social relations of production which better accommodate the expansion of society's productive capacity).

Since the above account, of how the growth of productive forces play a dominant role in Marx's economic analysis, is known as *historical materialism*, we must also explicitly include this materialist conception of history in the hard core of the Marxist research programme.[2]

The above discussion indicates the distinctive ways in which the hard core of the Marxist research programme differs, in our view, from the hard core of the orthodox research programme. It should be noted that the above discussion does not attempt to provide an exhaustive characterization of the hard cores of the orthodox and Marxist economics research programmes. Nor does it attempt to enumerate the different weights which various groups of orthodox and Marxist economists would give to each aspect of these hard cores. Rather, it attempts to identify the major non-vague features of these hard cores and to highlight their distinctive differences by contrast with one another.

4.3 *The Connection Between the Positive Heuristic and the Hard Core of a Research Programme*

Before proceeding to outline the positive heuristics of the orthodox and Marxist economics research programmes, recall that the positive heuristic of a research programme provides a set of suggestions as to how the hard core of the programme is to be supplemented in order for the programme to be capable of providing explanations and predictions of the phenomena under consideration.

---

[2] For support for the view that the productive forces have the above dominant role, as opposed to the view that the social relations of production are primary or the view that the forces and relations are mutually determining, see Cohen [1978] and Shaw [1978]. For an alternative view see Levine and Wright [1980].

As such, the positive heuristic guides the production of specific theories within the programme, with each specific theory not only being 'constructed around' the programme's hard core (or metaphysics) but also implying that hard core (or metaphysics). In other words, the distinctive metaphysics (or world picture) underlying a research programme will have an *influence* on theory construction within the research programme in the sense that each specific theory will not only be *built around* but will also *imply* the fundamental set of statements constituting this distinctive metaphysics (see Worrall [1978], p. 59). Also, in addition to guiding the production of specific theories within the programme, the positive heuristic provides guidance as to how to proceed when difficulties arise in the development of the programme.

4.4 *The Positive Heuristic of the Orthodox Economics Research Programme*

The positive heuristic of the orthodox research programme is viewed as containing guidelines which economics researchers may use in the various areas of *both* microeconomic and macroeconomic research activity. In other words, we view the positive heuristic of the orthodox research programme as guiding researchers in both the microeconomic and macroeconomic subprogrammes to make considerable use of identical analytical techniques and methods of solving problems. Thus, not only do both microeconomics and macroeconomics make extensive use of *equilibrium analysis, comparative-static analysis* (involving the displacement of an initial equilibrium to obtain a new or final equilibrium and then comparing the initial and final equilibrium situations), and *dynamic analysis* (involving an investigation of the path that the relevant economic variables follow when moving from the initial to the final equilibrium), but also in various areas of both microeconomics and macroeconomics we find that a *perfectly competitive market structure* has been assumed (meaning that the number of buyers and sellers in a given market is assumed to be large enough so that no one of them has any appreciable influence on the market price). Also, as researchers in macroeconomics have investigated the microeconomic foundations of their macroeconomic theories, considerable use has been made of *optimization analysis* in macroeconomics as well as in microeconomics. Moreover, researchers in both microeconomics and macroeconomics analyse the respective economic situations under consideration by constructing *a series of ever more complicated theories.*

Note, however, that the above view does not deny that the positive heuristic of the orthodox research programme has certain guidelines for macroeconomics research that are distinctly different (at least in emphasis) from those for microeconomics research. Hence, for example, more emphasis may be put on working at an aggregate rather than a disaggregate level, on short-run analysis rather than long-run analysis, and on output adjustment rather than price adjustment. Moreover, *within* both the microeconomic and macroeconomic subprogrammes of the orthodox research programme, even more specific areas of research (which we shall also refer to as subprogrammes in order to stress not only the programmatic aspect of the research activity but also that this research activity is part of the wider orthodox research programme) may also be identified by their distinctive positive heuristic guidelines (though this need not concern us here).

4.5 *The Positive Heuristic of the Marxist Economics Research Programme*

The distinctive differences between the hard core of the Marxist research programme and the hard core of the orthodox research programme indicate how Marx's economic analysis of mature capitalism stresses the explanatory roles of the forces and relations of production which characterize this specific stage of historical development. Hence, the positive heuristic of the Marxist research programme may be expressed generally as: 'explain how the historically specific set of forces and social relations of production, associated with the capitalist mode of production, determines various economic phenomena such as exchange ratios between goods, the distribution of total income between different classes, and the process of economic development'.

Note, however, that whenever the positive heuristic of the Marxist research programme provides a set of suggestions as to how the hard core of the programme is to be supplemented in order to provide an explanation of the various aspects of mature capitalism, these suggestions must ensure that the resulting analytical techniques, economic concepts and theories embody the essential features of the hard core hypotheses. Thus, for example, since Marx's economic analysis of capitalism stresses the historically specific set of social relationships between economic agents which characterize mature capitalism, then the positive heuristic of the Marxist research programme will require that each Marxist economic concept relates to this particular set of social or class relationships. In other words, the positive heuristic directs that the economic concepts of the Marxist research programme embody both a specific social and historical dimension.

To clarify the meaning of the preceding paragraph, we shall now provide a few brief illustrations of how analytical techniques and economic concepts embody a multi-dimensional character in Marxist analysis.

As a first example, note that although Marxist analysis includes equilibrium analysis and the method of comparative statics, these analytical techniques are applied in a way that always emphasizes the underlying historically specific set of social relationships. Hence, in Marx's use of these techniques, the specific structure, and changes in this structure, of social relationships is of particular importance. In contrast, the equilibrium analysis and comparative statics of the orthodox research programme do not have this sociological dimension. Moreover, while an equilibrium in orthodox economics is a position which is characterized by a balance of forces, and hence is a situation in which there is no tendency for change, the Marxist dialectical analysis emphasizes the hidden tensions or contradictions (essentially resulting from the underlying class struggle) within apparently static situations, which operate as a driving force leading to the emergence of a new situation.

As a second example, note that, in common with the orthodox research programme, the Marxist analysis commences with simplifying assumptions which, as the research programme develops, are subsequently replaced by more sophisticated and more realistic assumptions. However, while (in our view) the orthodox economics approach is to move from, for example, unrealistic idealizations of market exchange towards a better explanation of the reality of market exchange, the Marxist economics approach is very different.

Thus, in the Marxist analysis, the approach is to start with the 'reality' of social relationships which lie beneath the 'appearance' of market exchange and then, by successively more sophisticated logical analysis of the social relationships, to increasingly approximate the 'appearance' in terms of the 'reality' so as to move towards a better explanation of what the 'appearance' actually is. Moreover, the *logical stages* in this process of successive approximation correspond to specific *historical stages* in the development of capitalism. Hence Marxist analysis explicitly uses what is known as a 'logical-historical method'. In addition, this integration of the logical analysis and historical development of capitalism is dialectical in character, with the contradictions inherent in one stage leading to the emergence of the next stage.

As a further example, note that the economic concept of 'capital', in the Marxist analysis, is not viewed simply as a 'thing' (namely, the physical means of production). Rather capital is viewed as a particular social production relation which, in turn, is associated with a particular historical stage of development of society. Thus, in contrast to the concept of capital in orthodox economics, in the Marxist analysis the concept of capital explicitly includes both a social and a historical dimension.

As a final example of the multi-dimensional character of analytical techniques and economic concepts in Marxist analysis, note that while Marx, in common with many areas of orthodox analysis, assumed free competition between capitalists, the social and historical dimension enters once more in the sense that his analysis of competition also stresses the social relation of capitalist to capitalist, and of capitalists to workers, within a specific historical context. (For further discussion of Marxist analysis see Blaug [1980], Bottomore [1983], and Howard and King [1985].)

5 THE REGULATIVE INFLUENCE OF METAPHYSICS ON THEORY CONSTRUCTION IN ECONOMICS RESEARCH PROGRAMMES

In Section 4.3 we noted that the positive heuristic guides the production of theories within a research programme, with each specific theory being both constructed around and implying the programme's hard core or metaphysics. Clearly, if the underlying metaphysical statements constituting an economics research programme's hard core are vague, then the underlying metaphysics will have little if any regulative influence on theory construction within the research programme. However, since the metaphysical statements constituting the respective hard cores of the orthodox and Marxist economics research programmes have been shown to be both non-vague and distinctly different, then the specific metaphysics underlying each of these research programmes will have an important regulative influence on theory construction both by suggesting a certain range of theoretical possibilities and by explicitly ruling out a range of other theoretical possibilities.

While the specific metaphysics of the orthodox and Marxist economics research programmes do not provide *detailed* guidance about the exact form that any theory within these respective programmes might take, the positive heuristic of each programme restricts the range of theoretical possibilities by ensuring that the analytical techniques, economic concepts and theories embody the essential features of the respective underlying metaphysics.

An important illustration of this regulative influence arises with regard to the explanation of the economic phenomenon of capital accumulation. While both Marxist and orthodox economics attempt to explain capital accumulation, the Marxist analysis contains a specific social and historical dimension whereas the orthodox analysis is asocial and ahistorical in character. Thus, in the Marxist analysis, the motivation of the capitalist, lying behind the Marxist assumption of a strong drive to accumulate capital, is analysed as arising out of the process of historical development prior to mature capitalism, and as such, is linked to a specific social form of production. Moreover, in the Marxist analysis, individual capitalists are viewed as being compelled to accumulate capital (independent of their subjective preferences), with this compulsion on individual capitalists operating via the mechanism of competition. In contrast, orthodox economics analyses capital accumulation in terms of a general economic motivation of individuals to maximize wealth, and to achieve a preferred time-pattern of consumption, subject to certain constraints which embody no specific social or historical dimension.

In order to further highlight the importance of this regulative influence, let us assume that economics researchers are using the orthodox research programme as a basis for investigating economic situations. As noted above, this means that the specific *theories* constructed within this research programme will both be built around and imply the 'orthodox' hard core (or conjectural world picture). In addition, note that the *theoretical predictions* which will be generated will also imply the 'orthodox' hard core. In other words, the whole investigation is essentially concerned with providing an 'orthodox' explanation of the phenomena under consideration.

Moreover, note that if the theories, which have not only been built around but also imply the specific metaphysics of the orthodox research programme, are taken as a basis for making economic policy recommendations, and if a particular objective (for example, such as a reduction in unemployment) is assumed, then logical deduction will indicate what policy measures should be undertaken relative to this objective, Hence, in such a situation, the specific metaphysics of the orthodox research programme will underlie and influence not only the theoretical predictions but also the *economic policy recommendations*.

Clearly, if we had alternatively assumed that economics researchers are using the Marxist research programme as a basis for investigating economic situations, then the whole investigation would instead be concerned with providing a 'Marxist' explanation of the phenomena under consideration and the specific metaphysics of the Marxist research programme would consequently underlie and influence both the theoretical predictions and the economic policy recommendations generated within this research programme.

This link between the non-vague specific metaphysics of these economics research programmes on the one hand and the theoretical predictions and economic policy recommendations on the other hand highlights the importance of the regulative influence of metaphysics in economics. Also, it directly contradicts Koertge's claim that the heuristic significance of metaphysics in social science has been vastly overrated ([1979a], p. 22).

Before leaving this section, we would like to express our disagreement with Koertge on a further point concerning the regulative or heuristic function of metaphysics in social science. To do so, note that Koertge has also argued that in order for metaphysics to serve as a heuristic guide for scientific theorizing 'there must be *some* reason for taking the theories we are led to by the metaphysics more seriously than those theories it excludes' ([1979a], p. 28, her italics). Having introduced this additional, very stringent requirement, she then concludes that metaphysics cannot provide much heuristic guidance in this respect because metaphysics 'can rarely render an untested theory plausible' since 'typically there are either no reasons (or very, very weak reasons) for believing the metaphysics in the first place' ([1979a], p. 30).

While we accept that it is extremely difficult to argue that the specific metaphysics underlying an economics research programme can satisfy Koertge's additional, demanding requirement, this difficulty does not mean that the specific metaphysics is therefore heuristically insignificant. Rather, given (as we have argued and illustrated above) that metaphysics *does* have an important regulative influence in economics and (following Agassi) that researchers *are* especially interested in theories which may substantially enhance the understanding of a specific metaphysics, it is clear that metaphysics has an important heuristic significance in economics which should be explicitly recognized.

6 THE IMPORTANCE OF METAPHYSICAL PLURALISM

As well as stressing that metaphysics should be used with discretion within social science, Koertge has also stressed the importance of metaphysical pluralism as a protection against the danger of entrenched metaphysical dogmatism (see Koertge [1979a], p. 30). Since a variety of metaphysical frameworks will lead to a variety of scientific hypotheses being proposed, this means that although metaphysical frameworks cannot be empirically falsified they can be exposed to rigorous criticism by comparing the empirical and explanatory success of the explanations arising out of alternative metaphysical frameworks (see also Koertge [1981], p. 18). Not only does Koertge argue that this metaphysical competition will help to avoid entrenched metaphysical dogmatism, but she has further argued that whenever a result survives both the severe empirical testing and the rigorous criticism from alternative metaphysical perspectives there is 'reason to believe that the result may be near to the truth' ([1981], p. 21).

We agree with Koertge's argument about the danger of entrenched metaphysical dogmatism and the importance of metaphysical competition in social science. Given this agreement, it is our contention that the adoption of the MSRP method of appraisal in economics would help to make economics researchers more aware of the importance of metaphysical competition and thereby encourage the desired metaphysical competition. This contention is based on the argument that the adoption of the MSRP method of appraisal in economics has important pedagogic value in that it would help to make economics researchers aware (1) of the fact that there is a distinctive metaphysics underlying each of the major research programmes in economics; and (2) of the explicit link between the specific metaphysics of the hard core of a research programme on the one hand, and the theoretical predictions and economic policy recommendations generated within this research programme on the other hand; and (3) of the importance of competition between research programmes for progress (see Lakatos [1970], p. 155) and thus of the importance of the metaphysical competition implicit in comparing explanations based upon the different metaphysics of competing research programmes.

However, while the use of MSRP in economics would hopefully encourage metaphysical competition, we recognize that there is no guarantee that this will be the outcome or that entrenched metaphysical dogmatism will be avoided. We recognize that metaphysical competition may be hindered to some extent by difficulties in comparing the content of economics research programmes. Also, we recognize that the use of MSRP in economics may not prevent entrenched metaphysical dogmatism in that MSRP does not advise economics researchers to work only in progressive programmes and abandon degenerating rival programmes (see Lakatos [1971]). Rather, MSRP explicitly recognizes that it may be rational for some economics researchers to continue working in a degenerating economics research programme in the hope that it may be able to stage a comeback. Despite these difficulties, and the well-known criticisms of MSRP, it is our contention that since MSRP gives metaphysics a central, internal explanatory role in scientific explanation (rather than some external-influence role), and stresses the importance of competition between research programmes for progress, the adoption of the MSRP method of appraisal in economics could make a useful contribution by stimulating the desired metaphysical competition.

As noted above, Koertge has argued that if there is convergence on explanatory theories, after rigorous empirical testing and intense metaphysical competition, then this 'may indicate that we are approaching a correct understanding of the world' ([1981], p. 23). In clarifying what she means by this, Koertge argues (contrary to Watkins [1958]) that one cannot regard the specific metaphysics of a research programme as well confirmed because the theories built around and implying this specific metaphysics have been empirically confirmed by a large number of rigorous empirical tests. She argues (see Koertge [1979a], p. 29) that this method of confirming a specific metaphysics (or hard core) is unsatisfactory since it is tantamout to allowing the empirical confirmation of theories to count in favour of the specific metaphysics while, at the same time, not allowing anything to count against it (given that the hard core is regarded as not falsifiable).

Koertge has also argued (following Wisdom [1963]) that while the specific metaphysics of one research programme may be criticized by showing that it is inconsistent with a highly confirmed scientific theory produced by a rival research programme, 'it is very hard to make this kind of scientific "refutation" of metaphysics stick' if the latter research programme does not provide 'an ultimate, fundamental description of reality' ([1979a], p. 13). As Wisdom has pointed out (see Wisdom [1963], p. 305), if the refuting theory is later falsified, then the foregoing 'refutation' is automatically removed. However, despite the difficulties associated with this kind of scientific 'refutation' of metaphysics, Koertge takes the view that when a well tested scientific theory conflicts with metaphysics, it is the metaphysics which should give way and be abandoned (see Koertge [1979a], p. 28).

Unfortunately, while this kind of scientific 'refutation' of metaphysics may have taken place in the 'natural sciences', as Koertge has claimed (see Koertge [1979a], p. 28), it is extremely difficult to obtain such 'refutation' in the current state of economics. Given the highly complex, changing subject matter of economics, economics researchers not only face the difficulty of investigating a relatively large number of relevant variables in each research situation, but also the difficulty that these variables (and their interrelations not only with each other but also with other variables) are changing over time. This complexity of economic systems (together with the associated institutional changes and the much greater difficulty, or virtual impossibility, of controlled experiments in economics) has meant, in turn, that (despite the heroic efforts of econometricians to improve empirical testing in economics) it is extremely difficult for economics researchers to achieve decisive empirical results. Consequently, economics (in contrast, for example, to physics) has an extreme paucity of theories which (after rigorous testing) have remained unrefuted for a long period of time and which consistently manifest an exceptionally high degree of predictive accuracy.

The fact that debates in economics over the theoretical predictions and economic policy recommendations produced by competing economics research programmes are therefore often accompanied by the empirical difficulties noted above (despite attempts to obtain decisive empirical results), provides an important reason for ensuring that all concerned are aware of the specific metaphysics underlying the theoretical predictions and economic policy recommendations of the respective research programmes. This is especially so when it is recognized that the specific metaphysics underlying the Marxist research programme has a specific 'moral and political suggestiveness' (with this terminology taken from Watkins [1958], p. 357) which is very different from that of the specific metaphysics underlying the orthodox research programme.

As argued above, the adoption of the MSRP method of appraisal in economics would serve as an aid towards increased awareness of the specific metaphysics underlying the theoretical predictions and economic policy recommendations of a given economics research programme. Also, the use of MSRP would encourage metaphysical competition which, in turn, would help economic policy recommendations but also for moral and political recommendations.

University of Ulster at Coleraine

REFERENCES

AGASSI, J. [1964]: 'The Nature of Scientific Problems and their Roots in Metaphysics', in M. Bunge (*ed.*), *The Critical Approach to Science and Philosophy*, pp. 189–211. New York: Free Press.

ARCHIBALD, G. C. [1979]: 'Method and Appraisal in Economics', *Philosophy of the Social Sciences*, 9, pp. 304–15.

AVINERI, S. [1968]: *The Social and Political Thought of Karl Marx*. Cambridge: Cambridge University Press.

BLAUG, M. [1980]: *A Methodological Appraisal of Marxian Economics*. Amsterdam: North-Holland.

BOTTOMORE, T. (*ed.*) [1983]: *A Dictionary of Marxist Thought*. Oxford: Basil Blackwell.

COHEN, G. A. [1978]: *Karl Marx's Theory of History: A Defence*. Oxford: Clarendon.

GLASS, J. C. and JOHNSON, W. [1988]: *Economics: Progression, Stagnation or Degeneration?*, to be published by Wheatsheaf in Autumn 1988.

HAHN, F. and HOLLIS, M. (*eds.*) [1979]: *Philosophy and Economic Theory*. Oxford: Oxford University Press.

HOWARD, M. C. and KING, J. E. [1985]: *The Political Economy of Marx*, 2nd ed. London: Longman.

KOERTGE, N. [1979a]: 'Does Social Science Really Need Metaphysics?', English translation (kindly provided by N. Koertge) of 'Braucht die Sozialwissenschaft wirklich Metaphysik' in H. Albert and K. H. Stapf (*eds.*), *Theorie und Erfahrung*, Stuttgart, Klett-Cotta, pp. 55–81. (Page references are to the English translation.)

KOERTGE, N. [1979b]: 'The Problem of Appraising Scientific Theories', in P. D. Asquith and H. E. Kyburg (*eds.*), *Current Research in Philosophy of Science*, pp. 228–51. East Lansing, Michigan: Philosophy of Science Association.

KOERTGE, N. [1981]: 'Ideology, Science and a Free Society', English translation (kindly provided by N. Koertge) of 'Ideologie, Wissenschaft und eine freie Gesellschaft', in H. P. Duerr (*ed.*), *Versuchungen Aufsatze zur Philosophie Paul Feyerabends, 2. Band*, pp. 95–115. Frankfurt: Suhrkamp Verlag. (Page references are to the English translation.)

LAKATOS, I. [1970]: 'Falsification and the Methodology of Scientific Research Programmes', in I. Lakatos and A. Musgrave (*eds.*), *Criticism and the Growth of Knowledge*, pp. 91–196. Cambridge: Cambridge University Press.

LAKATOS, I. [1971]: 'Replies to Critics', in R. C. Buck and R. S. Cohen (*eds.*), *Boston Studies in the Philosophy of Science*, pp. 174–82. Dordrecht, Holland: Reidel.

LAWSON, T. [1985]: 'Uncertainty and Economic Analysis', *Economic Journal*, **95**, pp. 909–27.

LEVINE, A. and WRIGHT, E. O. [1980]: 'Rationality and Class Struggle', *New Left Review*, **123**, pp. 47–68.

SHAW, W. H. [1978]: *Marx's Theory of History*. London: Hutchinson.

WATKINS, J. W. N. [1958]: 'Confirmable and Influential Metaphysics', *Mind*, **67**, pp. 344–65.

WATKINS, J. W. N. [1975]: 'Metaphysics and the Advancement of Science', *The British Journal for the Philosophy of Science*, **26**, pp. 91–121.

WISDOM, J. O. [1963]: 'The Refutability of "Irrefutable Laws" ', *The British Journal for the Philosophy of Science*, **13**, pp. 303–6.

WORRALL, J. [1978]: 'The Ways in Which the Methodology of Scientific Research Programmes Improves on Popper's Methodology', in G. Radnitzky and G. Anderson (*eds.*), *Progress and Rationality in Science*, pp. 45–70. Dordrecht-Holland: Reidel.

Brit. J. Phil. Sci. 39 (1988), 331–352 *Printed in Great Britain*

Remarks on the Modularity of Mind

BENNY SHANON

ABSTRACT

Jerry Fodor's *The Modularity of Mind* is discussed. In this book the concept of modularity of cognitive processes is introduced and a picture of mind is proposed according to which the peripheral input systems are modular whereas the central processes are not. The present paper examines this view from both a methodological and a substantive perspective. Methodologically, a contrast between considerations of principle and of fact is made and implications for the nature of cognitive theory are discussed. Substantively, constraints on information flow are examined as they appear in various aspects of psychological phenomenology, and central processes in particular. It is suggested that the notion of modularity as *structural* and fixed be replaced by one which is *dynamic*, context-dependent. This modification, it is argued, is productive for the characterization of the workings of the mind, and it defines new questions for investigation.

1 *Introduction*
2 *Preliminary Remarks*
3 *Methodological Considerations*
4 *The Modularity of Perception*
5 *Modularity in Production*
6 *Central Modularity*
7 *Appraisal and Further Ramifications*

I INTRODUCTION

In his book *The Modularity of Mind* [1983] (henceforth, MoM) Jerry Fodor presents an overall picture of the cognitive system. In the main, this system consists of two distinct components: one is responsible for input analysis and is domain-specific and informationally encapsulated, while the other is responsible for the fixation of belief and is neither domain-specific nor encapsulated. Thus, the periphery, which is concerned with perception and language, is modular, whereas the center, which is concerned with what may loosely be characterized as thinking, is not. Graphically, this picture of the mind is shown by scheme (a) in Figure 1.

Scheme (a) defines a particular configuration in a two-dimensional space.

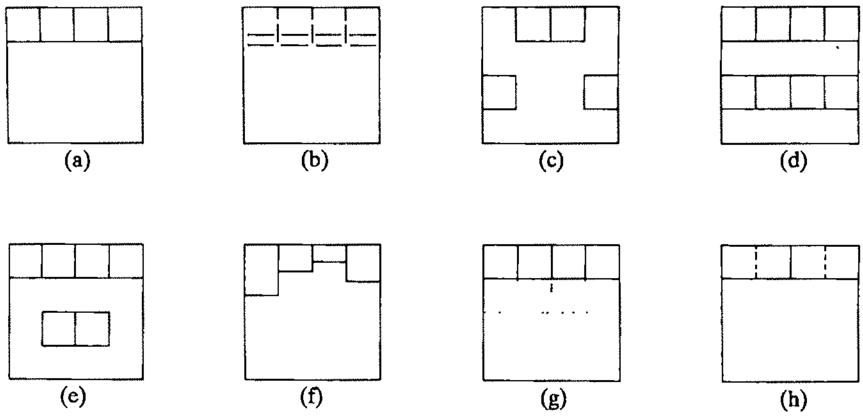


FIGURE 1. Various configurations of modularity.

The vertical axis specifies time: sections close to the top represent components responsible for early psychological processes, whereas the lower sections represent components responsible for later processes. The dimension represented in the scheme by the horizontal axis marks different aspects of psychological phenomenology.

The following discussion will examine several issues that are raised by MoM. It will examine methodological issues, explore aspects of mind that are not included in the MoM account, and will consider the modularity of central processes. On the basis of both substantive and methodological considerations an alternative view of modularity will be suggested. Consequently, more general implications regarding the structure of the mind and the prospects of its scientific investigation will be reviewed.

2 PRELIMINARY REMARKS

Modularity is defined in MoM by a profile of several characteristics, not all of which are on a par. These characteristics can be divided into three types: functional, procedural, and substrative. The functional characteristics define the logic of operation of the system under consideration: Modular components of mind are defined by domain-specificity, limited central access, and most importantly, informational encapsulation. The procedural characteristics specify the way the system actually operates in time: Modularity is associated with biological hardwiring and innateness as well as with well-defined sequences of development and patterns of neuropsychological breakdown. These three types of characteristics, I think, define a decreasing order of criteriality. Indeed, Fodor notes explicitly that the most important characteristics of modularity are domain-specificity and encapsulation, and that actually it is the last characteristic 'that makes a system modular'.

Given this notion of modularity the configuration shown in scheme (a) is defined. This configuration marks a series of specifications regarding the architecture of mind and its mode of operation. By way of setting the grounds for the discussion let me start by clarifying what configuration (a) specifies.

a. First and foremost, the configuration defines a particular partitioning in the structure of the mind. It differentiates between modular and non-modular components, and specifies the demarcation line between them.

b. The configuration assumes that different criteria of modularity all define divisions that converge upon the same line of demarcation. On the one hand, this line marks the end of perceptual processes; on the other hand, it marks the beginning of the processes of interpretation, comprehension and the fixation of belief. That the two demarcations converge upon the same psychological locus is, of course, not a logical necessity: Comprehension could have started before or after perception is achieved. The configuration of MoM thus imposes a significant constraint on the cognitive system, an alternative state of affairs is shown in scheme (b) of Figure 1.

c. According to scheme (a) exactly one line divides between the modular and non-modular components of mind. All components that are above that line (*i.e.* early in the course of processing) are of the first kind, whereas all components that are below it (*i.e.* later in time) are of the second. This specification excludes configurations such as the ones presented by schemes (c) and (d) of Figure 1. Like the configuration presented by (a) these two also differentiate between modular and non-modular components but they allow for modular components to be preceded by non-modular ones (as in scheme (c)) and to apply even after non-modular components have already acted (as in scheme (d), in which two demarcation lines between the modular and the non-modular are specified). As shown in scheme (e), the two options may be combined so as to define a configuration in which, following the application of peripheral modular components, some processes may apply as modular islands in an otherwise non-modular cognitive region.

d. The line is also one and only one in that it cuts across the different modalities in the same fashion. In this respect the scheme excludes a configuration like the one presented by scheme (f) in Figure 1, in which not all modalities are modular in the same fashion.

e. The configuration is fixed and invariant with time. It could be, however, that the property of being modular is task-dependent and that the particular location of the modular/nonmodular demarcation varies in different psychological contexts. Scheme (g) presents such a case.

f. Scheme (a) is read in one direction, from top to bottom or from without to within. Specifically, the cognitive perspective taken in MoM is that in which information arrives from the external world and is received by the human agent and then processed so as to eventually be interpreted and stored, thereby allowing for drawing conclusions and making inferences. There is, however, the converse course: the one that proceeds from within to without. This course presents itself in motor performance and in the production of language, as well as in processes of recollection, retrieval and problem-solving.

g. The configuration at hand also specifies a vertical demarcation. The input systems may be modular not only with respect to the central system but also with respect to one another. The two lines of demarcation are distinct. The input systems could be horizontally modular in the manner specified by scheme (a), yet exhibit vertical non-modularity of the kind graphically shown in scheme (h).

h. While Fodor mentions gradation, the picture of mind he presents does not leave room for such a variation. But, then, modularity could be conceived as being graded; for example, scheme (a) could be modified so that as one goes down the vertical axis the modular components become less and less distinct.

Given that distinct organs of our bodies specialize in the detection of different stimuli, it is necessary for there to be a level of the human information processing system that is modular. This level may be thin but it is bound to be there. Likewise, the non-modularity of at least some central processes cannot be dispensed with either. As Fodor points out, for an organism to function it should be able to integrate information even if this comes from different modalities, and it also has to coordinate between inputs and outputs. Having an identity of one's self calls for the same conclusion: At least some central core cannot be modular.

Taken together, these observations regarding modularity and non-modularity indicate that the issue of interest is not whether the basic pattern characterizing peripheral systems as modular and central ones as not modular does, in fact, hold. What is at stake, rather, is the particular form in which this basic configuration is instantiated and the further specifications that one may attribute to it. Indeed, the picture of mind presented in MoM is not a mere topological configuration. As noted above, Fodor emends the configuration he proposes with a series of particular specifications that he substantiates by reference to empirical data. Evidently, such a characterization of the modularity of mind is not one that could have been suggested on mere *a priori* grounds. In the main, the present discussion will not center on the particular empirical details that Fodor brings to support his view. It will focus rather on more general issues raised by MoM. The next section will focus on some methodological issues whereas the sections that follow it will examine substantive ones.

3 METHODOLOGICAL CONSIDERATIONS

A fundamental methodological question is that of evidence: What constitutes a proof that a system of mind is indeed modular? It seems to me that in this respect Fodor applies a double standard. In his discussion of the input systems Fodor is maximally rigid: By alluding to various auxiliary considerations he dismisses apparent instances of non-encapsulation as insufficient to indicate non-modularity. By contrast, in his discussion of the central processes he is very lax: It is enough for him to show that non-encapsulation is generally possible in order to conclude that the system in question is, indeed, not modular. In other words, for Fodor a few examples that an input system is not encapsulated are not sufficient for characterizing it as non-modular, whereas any instance of even potential nonencapsulation is sufficient for the characterization of the central processes as non-modular. The asymmetry of Fodor's argumentation may be further appreciated by the following considerations. Let us assume a cognitive function that enables the organism to modify the modularity of the components of the mind. Modular components may be locally made non-modular (temporarily influenced by the central system) and non-modular ones may be locally segregated and thus made modular. By Fodor's reasoning both components will be deemed non-modular.

Let me examine the two cases in turn. In discussing the informational encapsulation of sensory and linguistic perception Fodor admits the involvement of top-down processes but he argues 'that there are several prima facie reasons for doubting that the computations input systems perform could have anything like unlimited access to high-level expectations or beliefs. Even if there are some perceptual mechanisms whose operations are extensively subject to feedback, there must be others which compute the structure of a percept largely, or perhaps solely, in isolation from background information.' Is the cup half-full or half-empty, one may ask.

Fodor's severity in the attribution of modularity is marked when he emphasizes that mere consideration of cognitive phenomenology is not sufficient for the determination of modularity. 'To demonstrate [top-down] interaction between perception and background information is not, in and of itself, tantamount to demonstrating the cognitive penetrability of input processes. To get the latter result, you need to show that the locus of the top-down effects is *internal* to the input system; that is, you need to show that the information fed back interacts with internal levels of input processing and not merely with the final results of such processing. The penetrability of a system is, by definition, its susceptibility to top-down effects at stages *prior* to its production of output.' Such a determination is not a simple matter. As Fodor notes, 'It cannot be settled on the basis of the mere products of the system, but also requires reference to the internal processes involved in its operation.' The problem of assessing the informational encapsulation of input systems is thus 'not independent of the problem of determining how such systems are individuated and what sorts of representations constitute their outputs'. This assessment, however, may not be independent of one's general theoretical characterization of the system, and thus it is not entirely an empirical matter.

For example, an interaction of the type just noted will receive a totally different characterization according to how the input system is defined. A broad definition of the system will lead to the characterization of the interaction as internal hence to the conclusion that the system is not modular, whereas a narrow definition will lead to the characterization of the interaction as external, hence to the opposite conclusion. As noted, the choice bears on one's general view of the system, on one's intuitions, on one's stance regarding the nature of explanation in psychology, and on where one's interests lie. These are interrelated considerations which cannot be settled by simple observation of empirical facts.

A second methodological issue is more fundamental; it pertains to the perspective from which the discussion in MoM is conducted. While in his discussion of the input systems Fodor requires an inspection of the actual psychological mechanisms involved, in the discussion of central non-modularity his arguments are based on considerations of principle. Fodor defines central non-modularity by the two properties of isotropy and Quineanity. 'Isotropy is, by definition, the property that a system has when [a system] can look at anything it knows.' A system is Quinean if 'any given hypothesis is sensitive to properties of the entire system'. These definitions are taken from a *philosophical* rather than a *psychological* perspective, specifically, they are grounded in considerations of principle, ones that might be irrefutable. Thus, consider the following characterization of isotropy: 'By saying that [a system] is isotropic, [one] mean[s] that the ... relevant facts... may be drawn from anywhere'; 'everything that the scientist knows is, in principle, relevant to determining what else he ought to believe.'

Thus, 'In principle, our botany constraints our astronomy, if only we could think of ways to make them connect... Isotropy is, by definition, the property that a system has when [a system] can look at anything it knows.' The characterization is remarkable: The point argued, as well as the general context of the discussion, is cognitive-psychological, but the substance of what is being said pertains to the scientist. It is, however, not at all evident that arguments about scientific argumentation also hold for ordinary thinking. Thus, even if in principle the foregoing characterization is true, in fact it is not usually the case that cognitive agents regard their botany as related to their astronomy and, as will be further noted below, even if both are stored in their memory system they need not be connected.

The example just noted suggests that the considerations of principle Fodor employs in MoM are philosophical to an extent that takes them outside the scope of psychology proper. The philosopher's (as well as the logician's and the linguist's) perspective of principle is atemporal; the psychologist, by contrast, cannot avoid the consideration of the actual, and for him their is nothing more essential than time. Just as the atemporal study of language is linguistics but not psycholinguistics, the atemporal study of mind is epistemology but not psychology. Surely, the study of scientific discovery need not be the paradigm for the study of standard, everyday human reasoning.

Fodor's argumentation in MoM follows the philosopher's course. For him if information can be made relevant then the mind is (a non-temporal *is*) unencapsulated. I would say the opposite: if information is (in a temporal sense) inaccessible (even though if at another point in time it is accessible, even if in principle it could have been made accessible), then there is a property in the functioning of the mind that is to a degree modular.

The term 'to a degree' is important, for it is symptomatic of the difference in perspective noted here. Even though Fodor concedes in MoM that there are always variations of degree, the essence of his argumentation does not at all bear on these variations and does not take them into account. Central processes, he demonstrates, are not modular in the full sense of the term (and he is right in that respect); hence, he concludes, they are not modular, and thus, he continues, they are not constrained and they are as orderless as a Sears catalog. This admission of only two options—either fixed well-defined structures or else a complete mess—leads him to the conclusion that only the modular parts of mind are amenable to scientific investigation; whatever is non-modular is not. This conclusion results, in turn, in the extreme appraisal that the scientific study of cognition is not possible.

To the more general claims regarding the study of cognition I will return towards the end of this discussion. Here, I will focus on the methodological stance. Instead of arguing on the basis of the ever-present principled possibility of non-encapsulation I propose that we appreciate the actual, non-principled possibility of encapsulation. This change in perspective entails a change in the rule of argumentation. Fodor's argument is based on the (probably true) claim that there is nothing that is *in principle* encapsulated, and thus any proof of local sheltering of central structures and processes would be dismissed by him as not bearing on the attribution of modularity. The basis for the discussion that follows is the converse one: the appreciation that despite the principled possibility marked by Fodor the flow of information in central processes is subject to constraints, both general and *ad hoc*, and that this state of affairs is of primary cognitive significance.

4 THE MODULARITY OF PERCEPTION

Basically, I accept Fodor's characterization of the input systems as modular. Yet, I would like to point out that even here the situation may not be as categorical and fixed as it may appear at first sight, and that in some contexts patterns of non-modularity in perception may be encountered.

First, the boundary between the sensory periphery and the central system need not coincide with the boundary of comprehension. On the one hand, people can comprehend the meaning of verbal utterances even when their syntactic structure is not—perhaps cannot be—determined (*cf.* Schank [1973]). On the other hand, an utterance may pass the perceptual system yet not be comprehended. One may, for instance, keep a phrase in memory in its verbatim form and only later come to realize what that utterance means. The following episode will illustrate this point. Before entering the water in a Parisian swimming pool I put my towel on a hook. There were many towels there and I told myself 'Le deuxieme à gauche' (the second one from the left). Getting out of the water I wondered where my towel is. I did not know, but in my mind's ear I recalled the (verbatim) French phrase I uttered to myself before entering the pool. It was only then that the intact phonological form was interpreted, and I realized where my towel was: on the second hook from the left.

Turning to the (vertical) demarcation between the different modules of the input system, we note the following patterns of non-modularity. First, a stimulation in one sensory modality may be associated with the experience of sensation along another modality. This is the pattern defined by synaesthesia (Marks, [1978]). It may even be that the percepts detected by one sense organ are experienced as having the modality standardly associated with another. People under the influence of psychodelic drugs report such experiences. Second, sensory experience may depend on an interaction between different modalities. The interactions between the senses of taste, smell and vision define some basic axioms of gastronomy.

Crossing the (horizontal) line of modularity we note that the distinction between the different perceptual modes may persist well after the fast automatic processes have been completed and even past the achievement of comprehension. Modality-specific memories are a case in point (think of a pickle). The standard characterization of the human information-processing system is that it is modality-specific only in the early, peripheral stages and that as time goes on and processing progresses and departs from the periphery the modality distinctions are lost and retainment is increasingly abstract (for a paradigmatic empirical demonstration of this pattern see Sachs [1967]). Yet, there are indications that suggest that this is not categorically the case. Recent naturalistic studies reveal that people have remarkably good memory and detection ability for semantically non-significant changes in surface structure such as variations between nouns and anaphoric pronouns (Keenan, MacWhinney and Mayhew [1977]; Kintsch and Bates [1977]; Hjelmquist [1984]). Remarkable memory is also noted with regard to non-significant physical characteristics, be they in the visual or the auditory modality. In the auditory mode, Geiselman and his associates discovered memory for identities of speakers, their voice and location (Geiselman and Belezza [1976], [1977]); similar findings were also noted by Craik and Kirsner [1974]. Visually, people appear to remember non-significant details such as the particular location of items on the printed page (Rothkopf [1971]) and their typeface (Kirsner [1974]). Further, Kolers and his associates (Kolers [1975], [1976]; Kolers and Ostry [1974]) showed that subjects read text faster when it is presented in a typography or orientation in which they had read it before than when it is in another.

Finally, central involvement in perception may be noted as well. Fodor cites the persistence of perceptual illusions and their defiance of knowledge. While this is, indeed, the common characteristic of illusions, in some cases knowledge is more relevant than in others. For example, in the old witch/young lady drawing prior knowledge and experience do, in fact, bias one's manner of viewing the perceptual input. Contestable, too, is Fodor's characterization of the automaticity and encapsulation of the blink. The wearers of contact lenses have, with difficulty I admit, learnt to master the blink. This brings us to the consideration of the efferent aspects of behavior, to which I now turn.

5 MODULARITY IN PRODUCTION

As noted above, MoM follows one directionality: the efferent course of processing from the periphery to the center. But, surely, there is the converse, efferent course of processing, the one encountered in production and expression. It appears that these output systems exhibit patterns of modularity analogous to the ones Fodor notes in association with the input systems.

The case of linguistic production is most similar to that of linguistic perception examined in MoM. Linguistic specificity is clearly marked on the phonological level, where specific developmental sequencing and age dependency are even more marked in the latter case than they are in the former. The futility of central, guided control is also apparent. I, for example, conduct most of my professional life and much of my personal life in English, but my accent is poor and I cannot do anything about it. Indeed, whereas my general fluency in this language has constantly improved over the years, my accent has not changed much since the early stages of its acquisition, at about the onset of puberty. This poor accent is, I should note, not due to phonetic insensitivity, for I definitely hear the variations in different English dialects as well as individual idiosyncrasies in pronunciation. Finally, linguistic production is tied to neural architecture just as linguistic perception is, and as attested to by studies of aphasia (as well as agraphia) it reveals characteristic patterns of breakdown (*cf.* Geschwind [1974]).

With specialized motor skills, domain-specificity is evident. The hands of the piano player, for example, are tuned to the distances between the keys so that he/she can play even without visual feedback; this tuning, as well as that for touch, is not readily transferable to the playing of another keyboard instrument, say the clavichord. The dependence of motor production on the specific medium of its execution is further manifested in typing. Remarkably, even for an accomplished typist it is very difficult to indicate the location on the keyboard of the letters constituting a work without actually moving his or her fingers, in which case the task is very easy. Motor skills are clearly tied to body architecture. Skills are performed with particular muscles, they are grounded in well-defined neuronal structures and they are associated with specific patterns of breakdown in the case of brain damage (*cf.* Brown, [1972]).

Patterns analogous to the ones noted above seem to be associated with the expression of emotions. Facial expressions are clearly tightly linked with particular bodily patterns. Indeed, they allow for practically no distinction between their sense and the specific manner in which they are expressed: Changing the musculature of a smile and its mode of expression results in a totally different smile altogether. The resistance of these expressions to central control as well as their automaticity are revealed when one attempts (often without success) to fake an expression or when one's face, eyes or voice counter one's words or expressed attitude and reveal what one would have liked to conceal.

The characteristics of modularity noted above in conjunction with production need not require actual motor performance. In the domain of language relevant cases are cognitive performance—e.g., counting, calculating, reciting—tied to one's native tongue even in the context of another language (or linguistic community), and even when one is completely fluent in the latter (Shanon [1984]). Domain specificity is also encountered in other specialized cognitive skills. Idiots savants, for example, exhibit striking accomplishments in only a very particular domain of behavior, and the remarkable memory performance of chess masters is specific to chess (DeGroot [1965]). But, then, the characteristics of modularity associated with these tasks will not be lessened when the performances in question are not actually finalized in observable execution. Such patterns naturally lead us to the consideration of central modularity, which is the topic of the next section.

6 CENTRAL MODULARITY

At first blush it seems that it makes no sense to speak of modularity in the central processes. Fodor gives enough good functional reasons why these processes should be nondenominational and unencapsulated. As clearly noted, in order to fix one's beliefs one should certainly be able to operate on whatever information one has access to regardless of which cognitive domain the information is drawn from. Yet, as noted in the methodological discussion above, Fodor's characterization of central modularity is based on considerations of principle and follows a perspective which may by characterized as philosophical rather than psychological. While in principle the system of central processes may indeed be isotropic and Quinean, in fact it is possible that it may not be. Specifically, the respect by which the said properties apply to the system may be context-dependent, and the system may thus exhibit patterns of local modularity. It is here that the contrast between the philosophical and psychological perspective is apparent. Whereas for Fodor informational encapsulation that is context-dependent and varies with time is demonstrative of non-modularity (for the information at hand *can* be made relevant), the perspective proposed here suggests, by contrast, that it be taken as indicative of possible modularity (for the information can be put in abeyance and disregarded). This I say not because I deny the patterns of non-modularity, but because what I deem important is the dynamics of mind, and this dynamics consists, *inter alia*, of the changing of the boundaries of modularity. Thus, in proposing the attribution of modularity to central processes I am not arguing against the (probably true) claim that anything can, in principle, be made unencapsulated. Rather, I propose to change perspective and instead of looking at the ever-present possibility of non-encapsulation consider the actual and potential patterns of encapsulation. The following discussion examines various cognitive phenomena so as to show that central processes do exhibit patterns of modularity. Later in the discussion it will be argued that this attribution of modularity is not only meaningful but also productive.

The patterns of central modularity to be presented will exemplify the key defining characteristics of modularity—domain specificity and informational encapsulation; for each of the two families of modular patterns will be noted. Let us start with domain specificity. Just as perception may be modulated according to modality so may memory be modulated according to the characteristics of the particular context of acquisition at hand. The analogy hinges upon the fact that in both cases structuring is imposed by parameters that are not semantic, *i.e.* by parameters that are not part of the content of the information in question. Two families of phenomena attest to this: non-semantic constraints on entry to the memory, and medium and content effects in information processing. With respect to the first family of phenomena two cases will be noted: First, the phenomenon of encoding specificity (Tulving and Thompson [1973]), *i.e.* the dependence of retrieval on the particular characteristics of the context in which the information was first acquired. Akin to this is the dependence of retrieval on one's re-experiencing the particular episode in which the relevant information was acquired. Second to be considered are holistic patterns in retrieval. I refer to the impossibility of retrieving a particular item of information without retrieving the entire whole of which it is a part, *e.g.* the need to reproduce an entire recitation or melody so as to recall (or verify) one of its phrases. While not identical, these phenomena demonstrate the same basic pattern, namely, the relative reduction in the flexibility of retrieval from memory due to aspects of time or modality. These generate in memory sheltered residues, local rigid conglomerations that result in the relative encapsulation of information.

Domain specificity is noted both in memory and in the execution of cognitive skills, as well as in reasoning and problem solving. Patterns pertaining to the first two domains have already been noted earlier, in the discussion of perception and production. As noted, these patterns may constitute special intellectual achievements (as in the case of chess) or they may reflect the idiosyncratic affinities or the particular experiences of the individual. They may become, for the given individual, mandatory and automatic, and at times they may exhibit a marked resistance to control. By definition, these specializations are not innate, but, as we all know with time, habits form second nature.

As for reasoning and problem solving, the basic phenomenon of relevance is the following. Problems which are formally isomorphic may vary drastically in difficulty according to the particular content in which they are phrased. This phenomenon was demonstrated by Wason, Johnson-Laird and their colleagues (Wason and Shapiro [1971]; Johnson-Laird, Legrenzi and Legrenzi [1972]) by testing people with a problem like the following.

Four cards are displayed:

E K 4 7

and given that if a card has a number on one side it has a letter on the other (and vice-versa) the problem is which cards to turn over so as to determine whether the following statement is true or false:

If a card has a vowel on one side it has an even number on the other side.

It appears that the problem is very difficult to solve. By contrast, if the cards display any of the following:

Manchester, Sheffield, Train, Car

and the statement to verify is the following:

Every time I go to Manchester I travel by train.

the task turns out to be very easy. Similar patterns were found with other syllogistic (Wilkins [1928], Staudemayer [1975]) and mathematical problems (Gonzalez and Kolers [1982]). The common pattern is that a problem may be very difficult when phrased formally or in terms of an unfamiliar content-domain but quite easy when phrased in terms of familiar entities.

The phenomena pertaining to the central encapsulation of information are non-connectivity and mental islands. Non-connectivity, and hence inconsistency and incoherence, are inevitable consequences of the size of the human belief system and of the process of its construction. Clearly, it is not possible for one to draw all the conclusions that follow from each item of information one entertains nor is it feasible to constantly check for all the implications which information acquired in one context may bear on all other items in one's memory. As originally pointed out by Kant in his postulation of the synthetic *a priori*, the drawing of necessary implications is a non-trivial cognitive feat, and so is the bringing together of information drawn from different sources.

At times inconsistency serves one's interests. Prejudices mark this clearly. Prejudices are beliefs that are sheltered from one's other beliefs, perhaps even from manifest evidence that is presented to one. The Nazi S.S. officer who commanded over Jewish inmates may have believed that Jews at not human yet at the same time have held the belief that only humans can talk. His medieval predecessor may have believed that the Jew has horns or tails even after having seen one.

Further local cognitive segmentations are encountered when one is concentrating on particular problems or tasks. Concentration involves *ad hoc* focalization on some information at the expense of disregarding other, at times crucially relevant, information. The case of set shows precisely this: the local encapsulation of a particular content, perspective or way of processing from one's general knowledge or cognitive abilities.

Similar patterns are also encountered with phenomena that lie in the realm of 'personality'. Attitudes and affective reactions may exhibit at times striking specificity: a person who treats all people in a characteristically kind manner may fail to do so where his mother is concerned. Patterns analogous to the ones noted above in conjunction with non-connectivity are actually at the focus of the post-Freudian psychoanalytic theory of object-relations (Fairbairn [1963]). This theory characterizes behavior in terms of the presentations of basic relations that one has developed to objects (in this context, persons and, in particular, aspects of persons) other than oneself: the nourishing mother, the domineering one, the wise father, the angry one, and the like. Neurotic patterns are then characterized by the non-integration of these different relations, and it is the goal of the psychoanalytic process to bring them together.

In the extreme case note the segregation of the split personality.

For a last example of central non-connectivity, I would like to turn to my own research, to the study of *thought sequences, i.e.* of trains of verbal-like phrases that spontaneously pass through one's head (Shanon [1981], [1983a]). Specifically, consider the following three sequences:

1. 0. Eating soup with noodles, the noodles swim away.
    1. One should invent an eating utensil that has holes for the soup to drop off so that the noodles stay on it.
    2. But there is a fork.

2. 0. Drinking coffee sweetened more than usual.
    1. Good.
    2. It's the sugar.
    3. The sugar with the coffee is good.
    4. It's like chocolate.
    5. They could have done this with chocolate.
    6. But there is such a thing—cocoa.

3. 0. Listening to a composition and unable to identify it for quite a while.
    1. [half-consciously] It reminds me of 'The Swan'.

2. Saint-Saens.
3. Yes, they said [on the radio] it would be Saint-Saens.

The pattern of interest revealed by these sequences is that entities of information may be stored in the cognitive system yet not be connected to one another. Thus, the cognizer may reach a mentally stored entity not by means of a memory path that connects it to another such entity, but rather by thought processes which do not make use of already established connections. Sequences (1), (2), and (3) demonstrate this pattern in the lexicon, general knowledge of the world, and episodic memory, respectively; as shown by these examples, the pattern is characterized by a feeling of creativity or insight that immediately dissipates. In sequence (1), for example, it is clear that the cognizer knows that 'forks' are 'eating utensils with holes', yet in the particular context at hand he arrives at this appreciation by reasoning, as if *de novo*. This state of affairs, note, marks the contrast between considerations of principle and of fact to which we have already alluded in the methodological discussion above. In principle, given that the two entities are in the head, so to speak, there is no reason that they should not be connected. *De facto*, however, each connection between them is not available as such, and functionally they are relatively encapsulated with respect to the other. To conclude, even though everything in the mind is potentially interrelated, the mind is not isotropic in the sense that a space is.

Mental islands are defined by three characteristics. First, they are associated with a marked degree of vividness, as if they were real, yet it is clear that they are not (real). Second, no truth value is attributed to their contents—either because they are mental activities (not carriers of information) or because they are hypothetical. Third, functionally they serve as a workbench which is set apart, and toward which one directs the resources of one's cognitive system at large. By way of example, let me again consider some thought sequences:

4. 1. Where does one get a thought?
 2. Perhaps someone will sell me one?!
 3. 'Half a kingdom for a thought'.
 4. Laughter.

5. 0. Sitting in the kitchen.
 1. Nowadays, people put their ovens up.
 2. What if we put our stove up?
 3. Then I'd need a ladder in order to cook.

These two sequences are example of two ways in which information may be locally sheltered. Sequence (4) is an example of what I (see Shanon, in press) call *enactment*: the running of an internal theater, so to speak; in this particular example the enactment consists of a conversational fragment (line 3). In enactment one locally constructs a scenario in one's mind: the scenario allows vivid experience in a manner standard thought does not, but despite this feeling of reality it is clear that the situation is not real. Effectively, a separate mental island is created which is encapsulated from one's general system of knowledge and belief. Sequence (5) is an example of a *Gedanken-experiment*: one creates entities in one's mind and manipulates them so as to examine hypothetical states of affairs. Gedanken-experiments may be regarded as instances of mental manipulation of objects as if they were real. Yet it is clear that the configurations generated in thought experiments are not to be taken as true until they are expressedly acknowledged as such, in which case they may very well be incorporated within one's belief system. Both cases noted, then, are characterized by a marked flavor of reality coupled with a clear understanding that the things are not real. The information in question is encapsulated and thereby is endowed with special status. This encapsulation is localized not only in content but also in time: It is kept for a short period and then erased.

While enactment and Gedanken-experiments may be special, they are particular kinds of a variation which itself is standard. I refer to the plurality of *epistemic statuses*. Like the differentiation of percepts according to their modality, the contents of one's mind are treated differently according to whether they are considered as beliefs, knowledge, hypotheses, enactments and the like (for further development of these ideas, see Shanon [1988]). Functionally, the *ad hoc* sheltering of information just noted is very similar to patterns encountered in the execution of tasks in the real world. When constructing an object or repairing it, it is well to concentrate all required materials and tools on a workbench, and to separate them from other items so as to be able to direct one's resources. Experiencing a state of affairs as quite real and at the same time maintaining the knowledge that it is, in fact, not real is a fundamental ability of human intelligence. Without it one could not reflect upon the future, entertain possibilities, reminisce about the past, or engage in fantasy.

The cases of central modularity considered so far were not associated with clear substrative patterns like the ones noted in the periphery. I am, however, not at all sure, as Fodor is, that the lack of neurological correlates to the thought processes is a matter of principle. While the data are scant, it does appear that in frontal lesions one encounters patterned deficits which may be characterized as the breakdown of particular subroutines of central processes (see, for example, Pribram and Luria [1973]). And then there is one case of non-peripheral modularity which is clearly grounded in one's physiology. I refer to sleeping and dreaming. The states of wakefulness and of sleep are clearly segregated. Whereas the experiences of yesterday are both felt to be continuous with, and deemed relevant to, those of today, the experiences of yesternight are not. Specifically, when I open my eyes in the morning, I find (and I surely expect to find) myself in a situation that is on the whole continuous with that in which I was on the previous night when I closed them. The experiences of the previous day are also relevant in supporting reasoning with myself and arguments with others that I may conduct on the following day or days. The experiences of different nights (and to a great extent, the different dreams of one night), or the experiences of nights and days may not be related in this fashion. Thus, the experiences of days are not mutually encapsulated, while the experiences of nights are encapsulated with respect to those of the day as well as amongst themselves. The relative encapsulation of the experiences of the night with respect to those of the day calls for some clarification. Dreams often reflect experiences that happened during the day and in this respect the experiences of the night may be said not to be encapsulated with respect to those of the day. Yet, the involvement of day experiences in the dreams of the night is definitely different from the involvement of the experiences of one day in that of another. The involvement is by way of residues, leftovers that are no longer part and parcel of the experiences of the day in general. It is not the case that the experiences of the day can be used for the control and direction of whatever happens at night, and clearly one cannot during wakefulness take as evidential information drawn from one's dreams. The almost automatic forgetting of dreams is, in this sense, of prime survival value. Turning to the procedural characteristics of modularity, we note patterns that suggest an analogy with those encountered in conjunction with the paradigmatic modular systems. One cannot help dreaming nor does one have any control over its occurence. As already noted, dreaming is also associated with well-defined substrative characteristics: it is associated with particular neurological structure, with characteristic patterns of breakdown, and with definite sequences of development, both ontogenetic and phylogenetic (see Dement [1972]).

The different phenomena noted above are clearly not all of the same stock. Not all of them exhibit all the characteristics of modularity defined by Fodor, and those that do exhibit them to different degrees. By way of summary these phenomena may be grouped into the following five (not necessarily clearly distinct) main types.

First, there are those instances of central modularity that are ingrained in the architecture of mind; these are the most similar to the modular systems encountered in the periphery. In this class are the modular patterns associated with dreaming. These patterns are not a matter of choice and they may even be associated with well-defined substrative characteristics. Although acquired, severe abnormalities—like the ones encountered in schizophrenia—may also be included in this category.

Second, there are instances that are neither necessary nor built-in, yet *de facto* they are as deeply ingrained as the instances of the previous category. Included in this class are vested beliefs, *idees fixes*, values, idiosyncratic habits turned second nature, and well-internalized skills that one has developed in the course of one's life. Unlike the modalities of perception, which are defined by physics and biology, making their particular identity the same for all humans, the particular instantiations of the present class differ with individuals.

Third, there are cases that result from the size of the cognitive system and the fact that one cannot check context of acquisition. In this class are included the patterns noted in conjunction with beliefs, ethical conduct and object-relations. The relative encapsulation of some of these may be eliminated, but this may require at times substantial effort, both cognitive and affective.

Fourth, there is the *ad hoc* sheltering of information associated with the execution of particular tasks. This may be imposed upon one, as in the case of set in problem-solving, or generated by one, as in the cases of enactment and Gedanken-experiments. The encapsulation of these is temporary: Upon completion of the task the information may be erased or be integrated within one's general belief system.

Finally, there are residues: items that happen to maintain—in various degrees—their raw form and not be fully integrated into one's repertory of knowledge and belief. Such patterns are those associated with memories which are specific with regard to context, situation, modality or form.

7 APPRAISAL AND FURTHER RAMIFICATIONS

Given that the different cases noted above do not exhibit all the characteristics of modularity specified in MoM, one may dismiss them by saying that they do not pertain to modularity at all. In other words, it can be argued that the sense I attribute to the notion of modularity is so different from that of Fodor's that my discussion bears very little on that in MoM. Furthermore, it can be argued that my use of the term modularity is so lax that it loses much of its interest.

I would like to reject both reservations. First, the issue of definition. While admittedly patterns of central modularity surveyed here may not be considered as modular by Fodor, they all exhibit the crucial defining properties of modularity he specifies. The patterns in question vary with time and context but they do exhibit domain-specificity and informational encapsulation. As noted throughout the discussion, some patterns exhibit other (secondary) characteristics of modularity as well. Furthermore, the adoption of the strict definition of modularity results in the missing of significant psychological generalizations. The five categories listed at the end of the previous section blend into each other so as to define a continuum, which suggests that one is dealing here with different cases of the same basic phenomenon. Thus, it would be quite artificial to regard as modular only patterns of one category and not other patterns which are psychologically very similar to them. As a matter of fact, rather than being regarded as marking a weakness, the differences between the cases and their variation with context and time may be viewed as a genuine psychological phenomenon in its own right. That the modularity of mind is not fixed, but rather dynamically modifiable, is surely a phenomenon of cognitive interest.

By way of exemplifying the rigidity of Fodor's perspective let me note his likening of the central system of mind to a Sears catalog,[1] hence his appraisal that it is not worthy of scientific investigation. Fodor searches for well-defined structures and when he does not discover them he concludes that there is only one big mess. Yet between fixed nicely formalizable structures and a Sears catalog (assuming for the sake of the discussion that it is a total mess, which in all likelihood it is not) there are intermediate states of affairs. One such possibility is the study of the dynamics of the cognitive system. Rather than being associated with fixed structures that do not vary with time, the regularities of cognition may involve the logic of the system's mode of operation as it is conditional on context. That no formal structural regularities are found by no means implies that there are no regularities *tout court*.

It is, indeed, a shift towards a dynamic perspective of modularity which is suggested by the foregoing discussion. The modularity noted in MoM is *structural*: it is built in the (perhaps biological) architecture of mind, it is permanent, it does not vary with time and context, and clearly it is exclusively beyond the individual's control. The modularities noted here are *dynamic*: they do change with time and with context, they may be acquired and even be locally imposed or eliminated. The dynamic characterization of modularity surely does not exclude the structural one (although the converse is, of course, the case); rather, it includes it as one specialized case. The dynamic interpretation of modularity thus involves a relaxation of the meaning of the term. The purpose of the following discussion is to suggest that adopting this interpretation actually results in some advantages that would otherwise be lost. In other words, the present use of the notion of modularity is not only meaningful but also productive and beneficial.

The evaluation of the two perspectives bears on issues more general than modularity proper. In his book, Fodor not only considers the structure of mind with respect to modularity, but also makes general comments regarding the nature of cognitive theory and the feasibility of a cognitive science. Following his discussion of the non-modularity of the central processes Fodor proposes 'a generalization: one which [he] fondly hope[s] will some day come to be known as "Fodor's first law of the nonexistence of cognitive science", namely: The more global (*e.g.* the more isotropic) a cognitive process, the less anybody understands it. Very global processes, like analogical reasoning, aren't understood at all'. In line with this law he suggests that just as there is no serious philosophy of scientific confirmation so there is no serious psychology of central processes.

---

[1] Curiously, in this picture the central component of the cognitive system is actually of the associative type suggested in neo-connectionism (Hinton and Anderson [1981], Rumelhart and McLelland [1986], McLelland and Rumelhart [1986]).

The impoverished prospect for the study of cognition implied by Fodor's perspective appears even more severe when it is considered in conjunction with the moral presented in his 'Language of Thought' [1975] (LoT, for short). In this book, Fodor notes that 'cognitive psychology is concerned with the transformation of representations', and he contrasts it with 'psychophysics [which is concerned] with the assignment of representations to physical displays'. Yet, the transformation of representations, I take it, pertains to the central part of mind, whereas psychophysics subsumes the study of the input systems. Thus, what constitutes (according to LoT) the genuine subject matter of cognition is (according to MoM) unstudiable, whereas those modular parts of mind which (according to MoM) are the only ones that lend themselves to veritable scientific investigation are (according to LoT) not at all cognitive and to a large extent outside the scope of representational psychology altogether.

The source of Fodor's pessimistic conclusion is, I think, two-fold. Methodologically, it is in line with the formalistic philosophical perspective which assumes that for a theory to be scientific it has to abide by well-defined formal constraints. The modular parts of mind allow for that kind of theory, the non-modular parts seem not to. Substantively, the pessimistic conclusion follows from the structural perspective, according to which cognition should be concerned with those aspects of mind that are genuine manifestations of its architecture, not ones that may be changed by a person's beliefs and interest, *i.e.* not ones that are cognitively penetrable. This emphasis on fixed structure is in line with Chomsky's characterization of cognition (and of the study of language in particular) as the investigation of the organs of mind (*cf.* Chomsky [1980], as well as Pylyshyn [1980]).

Taken together, the foregoing considerations suggest that the adoption of the dynamic perspective is actually constructive and promising. Not only does it abrogate the verdict of impoverished cognition, but it may also open the door for the very psychology of thinking whose non-existence Fodor (in his less pessimistic appraisal) deplores. Whereas the structural perspective sets the psychologist to study the fixed, well-defined structures of the mind, the dynamic one calls for the study of the patterns by which these structures change. In the case of modularity this involves the processes by which informational encapsulation is achieved, maintained and eliminated, as well as the contexts associated with these processes, their benefits, and their costs.

By way of conclusion, let me summarize—in perhaps a somewhat speculative fashion—the picture of mind suggested by the foregoing discussion. As proposed here, modularity should be viewed not as a fixed structural property of mental organology but rather as a fundamental configuration that characterizes the dynamics of information processing in the mind. Functionally, this configuration seems to appear when the self is confronted with large bodies of information. Thus, in order to enable the organism to manage the perceptual information which is always there, aspects of modularity have evolved which are ingrained in the structure of mind and only in extreme contexts may be altered. These are the modularities associated with the input systems and which are presented in MoM. As the cognitive system develops and grows, one's own memory may potentially bombard one with information no less than perception does. Thus, in the interface between the self and its repertory of knowledge, *i.e.* in the every-present scanning of information in memory, patterns analogous to those encountered in perception are expected to be found. The size of one's system of knowledge and belief also results in segregations which at times (as in the case of failure to retrieve information from memory, in inconsistencies, in prejudicial behavior) are not necessarily to the advantage of the organism. There are also patterns which are context-dependent. When one focuses on the solution of a problem or on the execution of a particular task one would like to free oneself from information that is temporarily irrelevant. These situations are accommodated by the cognitive system's ability to generate its own modularities. These modularities lack the properties associated with fixed structures: innateness, neurological fixation, and specific patterns of development, nor are they necessarily associated with fast, automatic, mandatory processing. Yet there seems to be a similarity between the temporary mental autonomy generated in this fashion and the state of affairs encountered in the paradigmatic case of modularity, the one associated with perception. Like the perceived world, which is certain in its permanency and is there to be acted upon, so is the sheltered information in working memory (as well as in the cases of enactment and Gedanken-experiments noted above): Locally it is regarded as real.

The dynamics of modularity defines a bi-directional pattern. On the one hand there are the processes that generate modularity. These are encountered in the focalization on particular problems and tasks, in the development of skills, in the possession of accomplished expertise and in the establishment of habits. But then there is also the converse dynamics: the one that leads to less encapsulation and more isotropy. This dynamics is encountered in the integration of information and in what is characterized in common parlance as 'becoming more open'. Specifically it is associated with freeing oneself from one's prejudices, with the experience of insight—be it momentary or long-lasting, with the appreciation of meaningful patterns in the realities of nature and of one's life, as well as with the demolishing of old ideologies and the construction of the new as they occur both in scientific discovery and in artistic creation. All in all these patterns suggest that taken from a dynamic perspective the notion of modularity is actually one which is very useful for the study of cognition, both peripheral and central.[2]

---

[2] For a general critique of the structural-representational framework and for suggestions for the study of cognition from a dynamic perspective the reader is refered to Shanon [1987]; some of the ideas suggested are consistent (although motivated by different lines of reasoning) with the new cognitive paradigm of neo-connectionism (McClelland and Rumelhart [1986], Rumelhart and McClelland [1986]).

ACKNOWLEDGEMENTS

The ideas presented in this paper were developed in part in the course of a seminar on the philosophy of cognition given at the Hebrew University in the fall term of 1983. I thank the participants in the seminar for helpful discussions and Naomi Goldblum for her help in the preparation of the manuscript as well as for her comments and criticism.

Department of Psychology
The Hebrew University, Jerusalem, Israel

REFERENCES

BROWN, J. W. [1972]: *Aphasia, apraxia and agnosia: Clinical and theoretical aspects.* Springfield: Charles C. Thomas.

CHOMSKY, N. [1980]: 'Rules and representations', *The Behavioral and Brain Sciences*, **3**, pp. 1–61.

CRAIK, F. M. and KIRSNER, K. [1974]: 'The effect of speaker's voice on word recognition', *Quarterly Journal of Experimental Psychology*, **26**, pp. 274–84.

DEGROOT, A. D. [1965]: *Thought and choice in chess.* The Hague: Mouton.

DEMENT, W. C. [1972]: *Some must watch while some must sleep.* Stanford, CA: Stanford Alumni Association.

FAIRBAIRN, W. D. [1963]: 'Synopsis of an object-relations theory of the personality', *International Journal of Psychoanalysis*, **4**, pp. 224–5.

FODOR, J. A. [1975]: *The language of thought.* New York: Thomas J. Crowell.

FODOR, J. A. [1983]: *The modularity of mind.* Cambridge, MA: MIT Press.

FODOR, J. A. [1986]: 'Information and association', in M. Brand and R. M. Harnish (*eds.*), *The representation of knowledge and belief.* Tuscon: The University of Arizona Press.

GEISELMAN, R. E. and BELEZZA, F. S. [1976]: 'Long-term memory for speaker's voice and source location', *Memory and Cognition*, **4**, pp. 483–5.

GEISELMAN, R. E. and BELEZZA, F. S. [1977]: 'Incidental retention of speaker's voice', *Memory and Cognition*, **5**, pp. 658–66.

GESCHWIND, N. [1974]: *Selected papers on language and the brain.* Dordrecht: Reidel.

GONZALES, E. G. and KOLERS, P. A. [1982]: 'Mental manipulation of arithmetic symbols', *Journal of Experimental Psychology: Learning, Memory and Cognition*, **8**, pp. 308–19.

HINTON, G. E. and ANDERSON, J. A. [1981]: *Parallel models of associative memory.* Hillsdale, N.J.: Lawrence Erlbaum.

JOHNSON-LAIRD, P. N., LEGRENZI, P. and LEGRENZI, M. S. [1972]: 'Reasoning and sense of reality', *British Journal of Psychology*, **63**, pp. 395–400.

HJELMQUIST, E. [1984]: 'Memory for conversations', *Discourse Processes*, **7**, pp. 321–36.

KEENAN, J. M., MACWHINNEY, B. and MAYHEW, D. [1977]: 'Pragmatics in memory: A study in natural conversation', *Journal of Verbal Learning and Verbal Behavior*, **16**, pp. 545–60.

KINTSCH, W. and BATES, E. [1977]: 'Recognition memory for statements from a classroom lecture', *Journal of Experimental Psychology: Human Learning and Memory*, **3**, pp. 150–8.

Kirsner, K. [1974]: 'Modality differences in recognition memory for words and their attributes', *Journal of Experimental Psychology*, **102**, pp. 579–84.

Kolers, P. [1975]: 'Specificity of operations in sentence recognition', *Cognitive Psychology*, **7**, pp. 289–306.

Kolers, P. [1976]: 'Reading a year later', *Journal of Experimental Psychology: Human Learning and Memory*, **2**, pp. 554–65.

Kolers, P. and Ostry, P. J. [1974]: 'Time course of loss of information regarding pattern and analyzing operations', *Journal of Verbal Learning Behavior*, **13**, pp. 559–612.

Marks, L. E. [1978]: *The unity of the senses*. New York: Academic Press.

McLelland, J. L. and Rumelhart, D. E. [1986]: *Parallel distributed memory vol. 2*. Cambridge: MIT Press.

Pribram, K. H. and Luria, A. R. [1973]: *Psychophysiology of the frontal lobes*. New York: Academic Press.

Pylyshyn, Z. W. [1980]: 'Computation and cognition: Issues in the foundation of cognitive science', *The Behavioral and Brain Sciences*, **3**, pp. 111–69.

Rothkopf, E. Z. [1971]: 'Incidental memory for location of information in text', *Journal of Verbal Learning and Verbal Behavior*, **10**, pp. 608–13.

Rumelhart, D. E. and McLelland, J. L. [1986]: *Parallel distributed memory, vol. 1*, Cambridge: MIT Press.

Sachs, J. D. S. [1967]: 'Recognition memory for syntactic and semantic aspects of connected discourse', *Perception and Psychophysics*, **2**, pp. 437–42.

Schank, R. C. [1973]: 'Identification of conceptualizations underlying natural language', in R. C. Schank and K. Colby (*eds.*), *Computer models of thought and language*. San Francisco: W. H. Freeman.

Shanon, B. [1981]: 'Thought sequences and the language of consciousness', Annual meeting of the Cognitive Science Society, Berkeley.

Shanon, B. [1983]: 'Thought sequences', The Rotman Center for Cognitive Science, Hebrew University, Jerusalem, Israel.

Shanon, B. [1984]: The polyglot mismatch and the monolingual tie, *New Ideas in Psychology*, **2**, pp. 75–9.

Shanon, B. [1987]: 'The place of representations in cognition', in D. N. Perkins, J. Bishop and J. Lockheed (*ed.*), *Thinking*. Hillsdale N.J.: Lawrence Erlbaum.

Shanon, B. [1988a]: 'The channels of thought', *Discourse Processes*, in press.

Shanon, B. [1988b]: 'Enactment and Gedanken-experiments', *Journal of Mental Imagery*, in press.

Staudenmayer, H. [1975]: 'Understanding conditional reasoning with meaningful propositions', In R. J. Falmagne (*ed.*), *Reasoning: Representation and process in children and adults*. Hillsdale N. J.: Erlbaum.

Tulving, E. and Thompson, P. M. [1973]: 'Encoding specificity and retrieval processes in episodic memory', *Psychological Review*, **80**, pp. 352–73.

Wason, P. C. and Shapiro, D. [1971]: 'Natural and contrived experience in a reasoning problem', *Quarterly Journal of Experimental Psychology*, **23**, pp. 63–71.

Wilkins, M. C. [1928]: 'The effect of changed material on the ability to do formal syllogistic reasoning', *Archives of Psychology* 16 (102).

Brit. J. Phil. Sci. 39 (1988), 353–360 *Printed in Great Britain*

Hypothesis Tests and Confidence Intervals in the Single Case

D. J. JOHNSTONE

1 *N–P Theory in Ordinary Practice*
2 *Inference is Precluded*
3 *The Confidence Concept*
4 *Giere's Argument*
5 *Hacking's Argument*

I N–P THEORY IN ORDINARY PRACTICE

During all the years after 1933 until his death in 1981, Neyman insisted more and more that the methods he developed with Pearson are methods for 'inductive behaviour' (decision), not inductive inference.[1] Yet despite Neyman's strictures, statisticians in practice, including some of Neyman's most committed advocates, interpret their confidence intervals, and the results in hypothesis tests, as statements of *inference*. Indeed, it is recognized generally that although Neyman–Pearson [N–P] theory is widely accepted for its mathematics, especially for its concept of 'power', statisticians in practice have continued to interpret their results evidentially. This is the view of both Kyburg (a philosopher) and Birnbaum (a statistician):

> ...many practising scientists (particularly in psychology and the social sciences) ... automatically apply the statistical tests developed for the purposes of 'inductive behavior' to the evaluation of the significance of experimental results. (Kyburg [1974], p. 24)
>
> An evidential interpretation has been associated with typical applications of tests in scientific research investigations in all periods of their use (which dates from 1710), without apparent discontinuity during the years following 1933 when the mathematical structure of the Neyman–Pearson theory became widely accepted as the new or improved mathematical basis for the theory of tests. (Birnbaum [1977], p. 31)

---

[1] Johnstone [1987a] supports the claims of de Finetti that Neyman's philosophical writings on the interpretation of tests and confidence intervals are equivocal, or at least confused.

Most clearly, statisticians in practice tend to interpret a 95% confidence interval, *i.e.* an interval $f(X)$ which includes the true parameter Θ with relative frequency (Carnap's probability$_2$) 95%, as a statement that the degree of certainty (Carnap's probability$_1$) that the true parameter lies in the interval $f(X)$ is ·95. This is well known. As early as 1939, Harold Jeffreys ([1939], p. 300) declared that statisticians in practice, including Neyman and Pearson, tend, perhaps subconsciously, to interpret *relative frequency* (probability$_2$) as *degree of certainty* (probability$_1$) in the single case; (*cf.* Johnstone [1986], p. 496). More recently, various others have made this same observation. Kyburg, for example:

> ... many statisticians, and essentially all psychologists, sociologists, political scientists, economists, biologists, ecologists, bacteriologists, pathologists, physicians, toxicologists, astronomers, anthropologists, *etc.*, cite significance levels (the smaller the more proudly) and confidence levels (the higher the more joyfully) as though they reflected a level of evidential support applicable to the instance at hand ... ([1974], p. 58)

2 INFERENCE IS PRECLUDED

A confidence interval is an interval $f(X)$ such that the 'confidence coefficient' (probability$_2$) that $f(X)$ includes the true parameter equals $1-\alpha$. If the probability$_2$ $1-\alpha$ is interpreted as a degree of certainty (probability$_1$) in the single case, then the confidence interval $f(X)$ with 'confidence coefficient' $1-\alpha$ entails that (i) the probability$_1$ that the true parameter Θ lies within the interval $f(X)$ is $1-\alpha$, and (ii) the probability$_1$ that the disjunction $\{\theta_1$ or θ_2 or $\theta_3 \ldots\}$ of all $\theta \in f(X)$ is true is $1-\alpha$. Thus, either way, the confidence interval $f(X)$ with 'confidence coefficient' $1-\alpha$ entails a statement of *inference*. However, the probability$_2$ $1-\alpha$ is not a probability$_1$ in the single case other than by direct inference, for which N–P theory provides neither mechanism nor sanction.

Moreover, Neyman ([1971], p. 13) denied (somewhat reluctantly) that the probability$_2$ $1-\alpha$ may be interpreted as a probability$_1$ in the single case. This was an issue, especially with Fisher. Fisher (*e.g.* [1956], p. 86) maintained that so long as the reference set is conditioned on the 'relevant' respects of the sample X, specifically the value $A(X)$ of any ancillary statistic A, then the level of significance α (and thus the 'confidence coefficient' $1-\alpha$) is a legitimate probability$_1$ in the single case.[2] But apparently Neyman and Pearson were not familiar with Fisher's logic. Indeed, Pearson made it clear that he did not appreciate the sense in general of Fisher's word 'relevant'. He wrote:

---

[2] This interpretation of Fisher follows Johnstone ([1987b], pp. 482–483). Fisher's remarks on relevant subsets, conditionality, and ancillary statistics (*e.g.* [1955], pp. 69–78; [1956], pp. 86–92) constitute the logical antecedents of more recent principles of ancillarity, and other forms of partial conditioning; *cf.* Berger and Wolpert ([1984], pp. 11–18).

> Professor Fisher's choice of reference set is based, I think, on his theory of information. Thus in writing of the 2×2 table he speaks of 'the reasonable principle that in testing the significance with a unique sample, we should compare it only with other possibilities in all relevant respects like that observed', and again, in the regression problem, he refers to 'a population of samples in all relevant respects like that observed'. The meaning of terms such as 'relevant' is not of course self-evident without definition, but such phrases form part of Fisher's general approach to estimation theory and the reference sets adopted in these two examples are made perfectly clear. ([1955], p. 205).

In Fisher's terms ([1958], pp. 263–4), the respect $R(X)$ of the sample X is a 'relevant' respect (with respect to the test statistic t) if $R(X)$ defines a peculiarly distributed (with respect to t) subset of samples, *i.e.* if $f(t|R(X)) \neq f(t)$. For example, the sample size n defines a peculiarly distributed subset of samples, as does the value of an ancillary statistic. Thus, Fisher's principle that the reference set must be conditioned on the 'relevant' respects of the sample is simply the principle that the reference set must be *stochastically homogeneous*, *i.e.* free from any peculiarly distributed subset. This is the fundamental principle of direct inference, dating back at least to the writings of Venn, and emphasized (in various formulations) in perhaps every notable contribution on the logic for direct inference. Yet apparently Neyman and Pearson were not convinced that any such principle can be successful. Regretably, they did not ever say why.

3 THE CONFIDENCE CONCEPT

Prior to his death in 1976, Allan Birnbaum came to the view that there is an intuitive concept of statistical evidence implicit in N–P theory, at least in its application. He formalized this concept, which he labelled the 'confidence concept', as follows:

> (Conf): A concept of statistical evidence is not plausible unless it finds 'strong evidence for H_2 as against H_1' with small probability (α) when H_1 is true, and with much larger probability ($1-\beta$) when H_2 is true. ([1977], p. 24)

Briefly, Birnbaum's confidence concept is that a measure of statistical *evidence* is not a logical measure unless it has good error frequencies (α,β). This is not as strong as the 'confidence concept' which is often implicit in practice, for in practice, measures of statistical evidence, *e.g.* a measure which says 'accept' or 'reject' the null hypothesis at $\alpha = 5\%$, are justified on the basis of their error frequencies *per se*, *i.e.* good error frequencies are considered not only necessary, but *entirely sufficient*. This stronger confidence concept is explicit in a paper by Clopper and Egon Pearson, published in 1934. Clopper and Pearson declared that their *personal* level of 'confidence' (probability$_1$) that a confidence interval $f(X)=(p_1,p_2)$ includes the true parameter p depends only on the error frequency ('confidence coefficient') $1-\alpha$ attached. They wrote:

> Our confidence that p lies within the interval (p_1,p_2) will depend upon the proportion of times that this prediction is correct in the long run of statistical experience, and this may be termed the *confidence coefficient.* ([1934], pp. 404–5)

Hacking outlined the strong 'confidence concept' in his highly regarded *Logic of Statistical Inference* which was published in 1965. His response at that time, and again in ([1973], p. 493) was that the 'confidence coefficient' (probability$_2$) $1-\alpha$ is a legitimate measure of confidence or support (probability$_1$) in the single case only under a 'principle of irrelevance':

> In Neyman's theory, if you make a trial on the set-up and get result E for which $f(E)=\Gamma$, then Γ is an appropriate 95% *confidence interval.* ... it has been thought that you can be 95% confident that Γ does include the true distribution. Neyman may never have said so,[3] but other writers seem to have computed confidence intervals just because they favoured this interpretation. ... A 95% confidence interval is really worthy of 95% confidence on the data—or, at least, is an interval supported to degree 0·95—only where the principle of irrelevance can be used. ([1965], pp. 159–60)

Hacking's 'principle of irrelevance' requires that each respect $R(X)$ of the sample X is an 'irrelevant' respect, at least as far as we know, in the sense explicated by Fisher, *i.e.* in the sense that $f(p\in\Gamma|R(X))=f(p\in\Gamma)=1-\alpha$. Thus, for Hacking, Neyman's 'confidence coefficient' is a legitimate confidence coefficient (probability$_1$) in the single case only if the sample X is free (as far as we know) from any respect $R(X)$ such that $f(p\in\Gamma|R(X))\neq 1-\alpha$. By corollary, if the sample X has a 'relevant' respect $R(X)$, *i.e.* a respect $R(X)$ such that $f(p\in\Gamma|R(X))\neq 1-\alpha$, where p is the true parameter, then the 'true' confidence coefficient is not $1-\alpha$, but $f(p\in\Gamma|R(X))$. The probability$_2$ $f(p\in\Gamma|R(X))$ is given by conditioning the reference set on the respect $R(X)$; hence Fisher's principle that the reference set must be conditioned on the 'relevant' respects of the sample. A strong 'confidence concept' would seem to imply or presume this principle, for if the reference set is not conditioned where necessary, the probability$_2$ $1-\alpha$ is not a sensible probability$_1$ in the single case.

4 GIERE'S ARGUMENT

For now more than fifteen years, the philosopher Ronald Giere has maintained that Neyman's theory for estimation (confidence intervals) is saved from the problem of the single case if the probability$_2$ $1-\alpha$ is interpreted (following Peirce, Cramer, Braithwaite and Popper) as the measure of a 'propensity'. Giere's suggestion is that a system of confidence intervals has a physical 'propensity' or 'disposition' to generate intervals which include the true parameter, and the probability$_2$ $1-\alpha$ is the measure (manifestation) of this 'propensity'. Moreover, the 'propensity' $1-\alpha$ is physically constant, *i.e.* it holds the same physically in each single case. Hence, $1-\alpha$ may be interpreted as a probability (probability$_1$) in the *single case*. Giere put this as follows:

---

[3] I. J. Good [1984] cites published correspondence co-authored by Neyman in which the level of significance (P-value) observed in an experiment is interpreted implicitly as the posterior probability of the null hypothesis.

> In recent years a number of philosophers have proposed a 'propensity' interpretation of probability as an objective alternative to the traditional frequency interpretation. The prevailing view is that this interpretation automatically eliminates the problem of the single case. To say, for example, that the probability of a six with a fair die is 1/6 is to say that a certain physical set-up has a propensity, a physical disposition, to turn up sixes. Since this disposition is presumed to be present on every trial of the set-up, there is no question but that a statement of the probability of a certain chance outcome occurring applies to individual trials of the chance experiment. ([1969], p. 382)
>
> ... causal propensities are to be understood as operating during individual trials of probabilistic systems. Thus when a probability value is assigned to an outcome, that value attaches directly to the particular trial—no reference to any set of trials is necessary. ([1976], p. 71)
>
> ... propensities are physical tendencies existing in causally indeterministic processes. On this interpretation, propensities are conceived as operative on individual trials of indeterministic processes. Thus the objection that objectivists cannot assign error probabilities to individual inferences is overcome. ([1976], p. 100)

Giere's theory is not well received. To begin with, it has been explained by Salmon ([1971], pp. 39–40), Levi ([1973], p. 528), Kyburg ([1974], p. 10) and others of equal repute, that Giere does not avert the problem of the single case. To the contrary, Giere has the problem (in each single case) of identifying the particular physical routine in operation, and each particular *routine* corresponds to a particular *reference set*, *viz.* the set of results which that routine will generate if repeated *ad infinitum*. Thus, the problem of the single case, which is essentially the problem of identifying the appropriate reference set, remains for Giere as the problem of identifying the particular routine in operation. Giere gives the impression that this is no problem at all. But he deals only with mixed ('randomized') tests. These are tests in which the routine in operation, *i.e.* the routine chosen by the randomizer, is seemingly clear. But in other tests in which there is an ancillary statistic, the routine in operation is obscure. Consider, for example, the Behrens–Fisher test, in which the statistic s_1/s_2 is ancillary. Rosenkrantz [(1977], p. 219) maintains that the routine in this test is simply to generate sample pairs (x_1,x_2) with no particular value $y=s_1/s_2$, *i.e.* the value $y=s_1/s_2$ is merely incidental. However, Basu ([1964], p. 10) maintains that the routine in the Behrens–Fisher test has conceptually two stages; the first stage is to generate the value $y=s_1/s_2$, and the second stage is to generate the sample pairs (x_1,x_2) such that s_1/s_2 has that value y. There is nothing in Giere's writing to resolve this clash of intuitions. Giere ([1976], p. 99) claimed that his theory relates generally to tests in which there is an ancillary statistic, but he failed to elaborate. Thus, at best, his theory is incomplete. Rosenkrantz is stronger in his criticism. His position ([1977], pp. 218–20) is that if there is an ancillary statistic, the reference set (Fisher) or routine (Giere) must be conditioned on the value of that statistic. But given that Giere's theory ratifies conditionality only in mixed tests, in which the ancillary 'statistic' is the routine chosen by the randomizer, that theory is deficient.

5 HACKING'S ARGUMENT

Hacking [1980] retracted much of his well-known, and widely accepted, criticism of N–P theory. Hacking now takes the position that N–P theory is an admirable theory for inference. His explanation is that *inference is conclusion*, and that a theory for inference is a program for generating conclusions which are *usually correct*. His authorities are the philosophers Peirce and Brown. Brown had the idea that inference is conclusion, not argument, *e.g.* '*S*, so *H*' is argument, but not inference; the inference is just the conclusion *H*. Peirce proposed that a good argument is one which tends 'for the most part' to generate correct conclusions. Thus, following Hacking, N–P theory is *a Peircean theory for Brownian inference.*

Peircean theory precludes any concern with the single case. Yet many theorists, notably Fisher (*e.g.* [1955], p. 74; [1956], pp. 103–4), have long explained that in science the single case is crucial. The arguments are well known and utterly convincing. For example, Good ([1956], pp. 253–4) explained that if each single case is treated mechanically as just one in an indefinite sequence, sooner or later there will be a result which is so bad (ridiculous, or harmful) that there is no further single case to deal with, *e.g.* at best, the statistician may have his commission withdrawn. Other references include Cox ([1958], pp. 187–8), Pratt ([1961], p. 165), Kyburg ([1961], p. 298) and ([1974], p. 62), Lindley ([1965], p. 68) and ([1971], p. 31), Smith ([1965], p. 471), Edwards ([1969], p. 1235) and ([1972], p. 6), Giere ([1976], pp. 68–9) and ([1977], p. 45), Jaynes ([1976], pp. 200, 247), Rosenkrantz ([1977], p. 209), and Barnard ([1979], pp. 185–6). Yet perhaps the most damaging criticism of the school following Peirce, Neyman and Pearson is that Pearson himself was not a committed adherent. To the contrary, Pearson ([1938], p. 239) endorsed Student's argument that the prospect of very few errors in the 'long run' is cold comfort if the sample (in a single test) is patently atypical. Perhaps Hacking is willing to give up any concern with the single case, but Pearson was not; *cf.* Johnstone [1987a], pp. 267–68. This is a strong rejoinder to those who accept the philosophy of Neyman *and* Pearson.

REFERENCES

BARNARD, G. A. [1979]: 'In Contradiction to J. Berkson's Dispraise: Conditional tests Can Be More Efficient', *Journal of Statistical Planning and Inference*, 3, pp. 181–7.

BASU, D. [1964]: 'Recovery of Ancillary Information', *Sankhya*, **26**, pp. 3–16.

BERGER, J. O. and WOLPERT, R. L. [1984]: *The Likelihood Principle*. Institute of Mathematical Statistics.

BIRNBAUM, A. [1977]: 'The Neyman–Pearson Theory as Decision Theory, and as Inference Theory; With a Criticism of the Lindley–Savage Argument for Bayesian Theory', *Synthese*, **36**, pp. 19–49.

CLOPPER, C. J. and PEARSON, E. S. [1934]: 'The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial', *Biometrika*, **26**, pp. 404–13.

COX, D. R. [1958]: *Planning of Experiments*. Wiley.

EDWARDS, A. W. F. [1969]: 'Statistical Methods in Scientific Inference', *Nature*, **222**, pp. 1233–7.

EDWARDS, A. W. F. [1972]: *Likelihood*. Cambridge University Press.

FISHER, R. A.[1955]: 'Statistical Methods and Scientific Induction', *Journal of the Royal Statistical Society*, B, **17**, pp. 69–78.

FISHER, R. A. [1956]: *Statistical Methods and Scientific Inference*. Third Edition 1973. Hafner.

FISHER, R. A. [1958]: 'The Nature of Probability', *Centennial Review*, **2**, pp. 263–74.

GIERE, R. N. [1969]: 'Bayesian Statistics and Biased Procedures', *Synthese*, **29**, pp. 371–87.

GIERE, R. N. [1976]: 'Empirical Probability, Objective Statistical Methods and Scientific Inquiry', in W. L. Harper and C. A. Hooker (*eds*), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Volume 2, pp. 63–101. D. Reidel.

GIERE, R. N. [1977]: 'Testing Versus Information Models of Statistical Inference', in R. G. Colodny (*ed.*), *Logic Laws and Life*, pp. 19–70. University of Pittsburgh.

GOOD, I. J. [1956]: 'Which Comes First, Probability or Statistics?' *Journal of the Institute of Actuaries*, **82**, pp. 249–55.

GOOD, I. J. [1984]: 'An Error by Neyman Noticed by Dickey', *Journal of Statistical Computation and Simulation*, **20**, pp. 159–60.

HACKING, I. [1965]: *Logic of Statistical Inference*. Cambridge University Press.

HACKING, I. [1973]: 'Propensities, Statistics and Inductive Logic', in P. Suppes, *et al.* (*eds*). Logic, Methodology and Philosophy of Science IV, pp. 485–500. North Holland.

HACKING, I. [1980]: 'The Theory of Inference: Neyman, Peirce and Braithwaite', in D. H. Mellor (*ed.*), Science, Belief and Behaviour: Essays in Honour of R.B. Braithwaite, pp. 141–60. Cambridge University Press.

JAYNES, E. T. [1976]: 'Confidence Intervals vs. Bayesian Intervals', in W. L. Harper and C. A. Hooker (*eds*), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Volume 2, pp. 175–257. D. Reidel.

JEFFREYS, H. [1939]: *Theory of Probability*. Oxford University Press.

JOHNSTONE, D. J. [1986]: 'Tests of Significance in Theory and Practice', *The Statistician*, **35**, pp. 491–504.

JOHNSTONE, D. J. [1987a]: 'On the Interpretation of Hypothesis Tests Following Neyman and Pearson', in R. Viertl (ed.), Probability and Bayesian Statistics, pp. 267–77. Plenum.

JOHNSTONE, D. J. [1987b]: 'Tests of Significance Following R. A. Fisher', *The British Journal for the Philosophy of Science*, **38**, pp. 481–99.

KYBURG, H. E. [1961]: *Probability and the Logic of Rational Belief*. Wesleyan University Press.

KYBURG, H. E. [1974]: *The Logical Foundations of Statistical Inference*. D. Reidel.

LEVI, I. [1973]: 'Book Review: The Matter of Chance by D. H. Mellor', *The Philosophical Review*, **82**, pp. 524–30.

LINDLEY, D. V. [1965]: *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Part 2. Cambridge University Press.

LINDLEY, D. V. [1971]: Discussion of 'When Is Inference Statistical Inference?' by M. S. Bartlett, in V. P. Godambe and D. A. Sprott (*eds*), *Foundations of Statistical Inference*, pp. 20–31. Holt, Rinehart and Winston.

NEYMAN, J. [1971]: 'Foundations of Behavioristic Statistics', in V. P. Godambe and D. A. Sprott (*eds*), *Foundations of Statistical Inference*, pp. 1–19. Holt, Rinehart and Winston.

PEARSON, E. S. [1938]: '"Student" as Statistician', *Biometrika*, **30**, pp. 210–50.

PEARSON, E. S. [1955]: 'Statistical Concepts in Their Relation to Reality', *Journal of the Royal Statistical Society*, B, **17**, pp. 204–7.

PRATT, J. W. [1961]: 'Review of Lehmann, E. L.: Testing Statistical Hypotheses', *Journal of the American Statistical Association*, **56**, pp. 163–7.

ROSENKRANTZ, R. D. [1977]: *Inference, Method and Decision*. D. Reidel.

SALMON, W. C. [1971]: *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press.

SMITH, C. A. B. [1965]: 'Personal Probability and Statistical Analysis', *Journal of the Royal Statistical Society*, A, **128**, pp. 469–99.

Brit. J. Phil. Sci. 39, (1988), 361–366 *Printed in Great Britain*

Azande Logic *Versus* Western Logic?

TIMM TRIPLETT

ABSTRACT

In *Knowledge and Social Imagery*, David Bloor suggests that logical reasoning is radically relativistic in the sense that there are incompatible ways of reasoning logically, and no culturally transcendent rules of correct logical inference exist which could allow for adjudication of these different ways of reasoning. Bloor cites an example of reasoning used by the Azande as an illustration of such logical relativism. A close analysis of this reasoning reveals that the Azande's logic is in fact impeccably Aristotelian. I argue that the conclusions Bloor can legitimately draw from his case study are not controversial and do nothing to make plausible the thesis of logical relativism.

The thesis that the knowledge we claim to possess is culturally relative in some philosophically significant sense has gained increasing prominence in recent years. Relativism has for some time been a concern of many sociologists of knowledge and philosophers of science.[1] More recently it has entered into the heart of debates about the purpose and future direction of philosophy as a whole. The moves toward 'post-analytic' philosophy and 'cultural hermeneutics' prominently include the assertion that knowledge must be viewed relativistically. The relativist claims not just that different cultures or historical eras will know different propositions, but also that there is no culturally transcendent concept of knowledge which will allow objective adjudication of a given knowledge-claim. Not only is a knowledge-claim sufficiently legitimized for a given culture if it is accepted as knowledge within that culture, there is in fact no higher epistemic status possible for a proposition.[2]

Differences arise within the relativist camp, however, regarding the scope of relativism. Knowledge of ethical or aesthetic propositions might be accepted as relative, but heels begin to dig in when it comes to the propositions of the natural sciences. And the claim that our knowledge of logical truths is relative is still beyond the pale for most philosophers.[3]

[1] Seminal works in these fields are Karl Mannheim [1936] and Thomas Kuhn [1962].

[2] For recent expressions of this sort of relativism see *e.g.* Richard Rorty [1979], pp. 170, 315–22, 373–94; John Rajchman [1985], pp. xiv–xvi; and Brice Wachterhauser [1986], pp. 7–8.

[3] Even sociologist of knowledge Mannheim, for example, explicitly excludes logic and mathematics from the scope of the sociology of knowledge. (See Mannheim [1936], p. 43.)

In *Knowledge and Social Imagery*, David Bloor asks the reasonable question why one should dig in one's heels at one point rather than another.[4] He concludes that there is no good reason why: in fact, our perception that knowledge of science, mathematics and logic is somehow less susceptible to cultural conditioning turns out itself to be a culturally conditioned perception. Bloor's 'strong programme in the sociology of knowledge' holds that sociology can explain the truth of a proposition just as readily as it can explain that proposition's social genesis or historical impact (pp. 4–5, 10–13). And Bloor emphatically includes mathematical and logical knowledge within the scope of his strong programme (Chs. 5–7, *passim*). My concern in this essay is with the alleged relativity of logical knowledge.[5]

Does Bloor really mean to claim that the fallacy of affirming the consequent might be correctly known to be a legitimate logical inference in some other culture—that it is only a fallacy in the West, or (perhaps) only in analytic philosophy departments? He certainly creates the impression that he is defending this sort of radical logical relativism. He claims that he will provide examples that illustrate how sociology can 'explain the logical necessity of a step in reasoning [and] why a proof is in fact a proof' (p. 74). If the explanation is sociological, then it would seem that what is logically necessary in one society need not be so in another. Bloor also states that the successful application of the strong programme should reveal methods of reasoning that 'violate our sense of logical and cognitive propriety' (pp. 95–6). He claims his approach is relativistic (p. 142), accepts objectivity only in a sense a radical relativist would be happy with—objectivity is 'institutionalised belief' (p. 87)—and implies throughout his book that his strong programme is a significant, radical and controversial innovation. Finally, he offers in the section I am concerned with a cultural study of the Azande in which he suggests that 'there must be more than one logic: an Azande logic and a Western logic' (p. 124).

But beneath this surface impression of a radical assault on the traditional concept of the objectivity of logical reasoning, matters are not so clear. Bloor will suggest logical relativism in one passage only to suggest a more modest conclusion—or an ambiguous one—in another. In part the very fact of such unclarity makes a careful examination of relativist arguments important. Vague premises and ambiguous conclusions can too easily create the impression that some sort of philosophically interesting relativism has been established or at least strongly argued for. I will argue that Bloor trades on ambiguity, that what is presented as radical and innovative is in fact true only in the least controversial of senses.

Bloor's study of the Azande's logical reasoning is based on data presented by

---

[4] David Bloor [1976]. All page references in the text are to this work.

[5] I discuss Bloor's sociology of mathematics in Triplett [1986].

Evans-Pritchard.[6] The Azande are represented, in Bloor's terms, as believing that witchcraft 'is an inherited physical trait, consisting of a substance in the belly called witchcraft-substance' and that 'a male witch will transmit witchcraft-substance to all his sons, and a female witch to all her daughters' (p. 123). Bloor then argues that 'it would seem a clear logical inference that only one, single, decisive and incontestable case of witchcraft is needed to establish that a whole line of people have been or will be witches' (p. 124). In fact, the Azande do not make this inference. Bloor suggests that this implicates logic in some serious way.

What precisely do the Azande conclude? Apparently, they regard it as impossible that a whole clan could be witches. They reject the allegedly inevitable logical conclusion that all a witch's same-gendered descendants are witches by making a distinction between a true witch and one who possesses the witchcraft-substance but is not a witch. The line the Azande take is that the witchcraft-substance does not invariably confer 'witchhood' on the possessor of that substance. In some cases the witchcraft-substance is 'cool' and its possessor is for all intents and purposes not a witch (see p. 126).

Here it begins to become unclear just what conclusion Bloor wants to draw from these facts about the Azande. He seems to be entertaining, and perhaps confusing, two distinct conclusions. On the one hand his point at times seems to be that logic is no threat to the Azande's social institutions, that logic itself is impotent to force a society to accept a conclusion it does not want to accept or cannot accept if it is to survive intact. (See, for example, the diagram illustrating 'the impotence of logic', p. 126). The second and more radical point Bloor hints at is logical relativism proper: there just are alternative and incompatible ways of reasoning logically and there are no culturally transcendent rules of correct logical inference that can allow us to comparatively evaluate these different methods of reasoning. This relativistic conclusion is suggested by the remark that Azande logic is different from Western logic (p. 124), together with the relativistic implications and statements noted above that occur throughout Bloor's book. I shall consider the relativistic conclusion first.

It is not clear from Bloor's account how the Azande originally construed the relation between being a witch and possessing the witchcraft-substance. Depending on how one does construe this relation, one can construct two different accounts of the reasoning used by the Azande. But, as I will show, on neither account is there anything to indicate that the Azande use a logic that is incompatible with ours.

Suppose, to consider one reading of this relation, that the Azande originally maintained only that every witch had the witchcraft-substance, and never addressed the question whether every person who had the witchcraft-substance was a witch. Consider then the following argument:

---

[6] E. E. Evans-Pritchard [1937]. The present discussion is based on data from this book as presented by Bloor.

(1) Every witch has the witchcraft-substance.
(2) The witchcraft-substance is always inherited by the same-sexed children of a witch.
(3) Male A and female B were both witches.
(4) Everyone in clan C is descended from A or B.

---

(5) Everyone in clan C is a witch.

If premises 1 and 2 are the basis of Azande witch lore, then the Azande are perfectly justified in rejecting conclusion 5, for it does not follow as a matter of Aristotelian logic. The first premise says only that every witch has the witchcraft-substance, not that everyone who has the witchcraft-substance is a witch. It would appear that, far from developing an alternative logic, the Azande are being particularly careful here not to draw conclusions that do not logically follow.

Suppose though that as a matter of fact it had always been assumed in Azande society, though perhaps not explicitly stated, that anyone possessing the witchcraft-substance was a witch, as might be reasonable given the other assumptions. Then we have in place of 1:

(1′) Every witch has the witchcraft-substance, and everyone possessing the witchcraft-substance is a witch.

Given this alternative premise and some additional obvious premises concerning the relations of inheriting and being a descendent, the conclusion 5 does logically follow from the Azande premises. Suppose now that a Western anthropologist enters the scene and draws attention to the fact that their premises seem to implicate all of clan C as witches. Does the fact that the Azande resist this conclusion show that they have an alternative logic, or that they reject logic as we understand it? Not at all. They are simply reasoning according to the method of *reductio ad absurdum*. Since the conclusion that all the members of the clan are witches is unacceptable, there must be something wrong with one or more of the premises. This is in fact precisely how the Azande respond according to Bloor's account: They reply that sometimes the witchcraft-substance is 'cool' and does not lead to its possessor being a witch, *i.e.* they deny premise 1′. Perhaps, if the Azande did initially believe something like 1′ as opposed to the weaker 1, they have been confronted for the first time with an inconsistency in their set of beliefs. But they do not then hang on to all their beliefs and maintain that it's fine to be inconsistent. They revise their beliefs so that they are consistent.

There is nothing here that is illogical about the Azande's reasoning. It is, in fact, quite Aristotelian. What is happening here is that the deflection of the allegedly logically inevitable conclusion is not an abandonment of logic but an abandonment of premises that logically lead to a conclusion that is, to the Azande, unacceptable. The conclusion's unacceptability is based on Azande factual beliefs about the likelihood of pervasive witchhood in their villages and has nothing to do with adopting an alternative and incompatible logic. Far from vindicating Bloor, the example of the Azande gives support to the idea of a universal logic that is, even if not articulated or explicitly studied, adhered to in all cultures in their practical reasoning. Indeed, as Bloor himself rather lamely points out (pp. 128–9), this process of elaboration and revision goes on all the time in Western science. Bloor even notes that 'the Azande think very much as we do' (p. 129). That noted, Bloor asks the necessary question: 'Where does this leave the question of whether the Azande have a different logic from us?'[7] Bloor concludes:

> The Azande have the same psychology as us but radically different institutions. If we relate logic to the psychology of reasoning we shall be inclined to say that they have the same logic; if we relate logic more closely to the institutional framework of thought then we shall incline to the view that the two cultures have different logics. It would accord with the previous chapters on mathematics to choose the latter course. Far more important than such definitional matters however is the basic acknowledgment that both psychological and institutional factors are involved in reasoning (pp. 129–30).

This conclusion is confusing and ambiguous. It is unclear what it is to 'relate logic to the psychology of reasoning' or to 'the institutional framework of thought'. The 'more important' conclusion that 'both psychological and institutional factors are involved in reasoning' is again unclear. Just how are they involved? We have seen above that there is no reason to assume that the 'involvement' is so strong as to make plausible the conclusion that the Azande operate according to some strange non-Western logic. Quite the contrary.

Bloor's argument, then, does nothing to support logical relativism. But what about Bloor's other suggestion that logic is impotent to force the Azande to challenge beliefs that are important to the maintenance of their social institutions? This suggestion implies something about logic and society that is correct but uncontroversial. The conclusions Bloor can legitimately come to on the basis of his case study of the Azande concern factual beliefs and hypotheses, not logical truths. The first of these conclusions is that a society's institutional structures, as well as the individual psychologies instilled in its members, can cause some factual beliefs to be held very strongly, so strongly that they will not be given up when found to logically conflict with other factual beliefs. Those other beliefs will be revised instead. (The fact of such revision indicates that the need for logical consistency is respected.) The second conclusion is that the Azande institutions are different from ours, so the factual beliefs which Azande society will hold most strongly are likely to be different from those which we hold most strongly. As far as the question of the radical relativity of logic goes, these conclusions actually *support* the anti-relativist idea that the Azande recognize the same concepts of logical consistency and validity as we do.

---

[7] Bloor had never prior to this point presented this as a 'question' to be decided. He had instead implied that he had already decided the issue in favor of the existence of two distinct logics and it only remained to illustrate this conclusion (see Bloor [1976], p. 124). This is one example of unclarity in the structure of Bloor's presentation.

I will not say that these two conclusions are trivial truths, for the recognition that social and psychological factors enter into people's willingness to accept and propagate various beliefs was an important achievement in the history of ideas. But these points are hardly news. Nor are they controversial. Few if any in this post-Marxist, post-Freudian world will deny that what is accepted and vehemently defended as fact can be strongly influenced by the nature of the society, class or individual making that defense, particularly when matters of ideology (personal or social) are at stake.

Bloor's attempt to write an interesting sociology of logic thus fails. He has not established any reason for thinking it plausible, let alone necessary, to posit incompatible and 'equally valid' alternatives to our logical reasoning and knowledge of logical truths. And where Bloor himself seems to recognize this and, sometimes confusedly, to turn to a weaker conclusion, that conclusion is merely a restatement of the well-accepted view that social institutions can create a powerful inclination to believe among that society's members. Such a conclusion, while not objectionable, is hardly such as to provide striking new insights or to legitimize an interestingly strong programme in the sociology of knowledge.

University of New Hampshire

REFERENCES

BLOOR, D. [1976]: *Knowledge and Social Imagery*. Routledge & Kegan Paul.

EVANS-PRITCHARD, E. [1937]: *Witchcraft, Oracles and Magic Among the Azande*. Clarendon Press.

KUHN, T. [1962]: *The Structure of Scientific Revolutions*. University of Chicago Press.

MANNHEIM, K. [1936]: *Ideology and Utopia*. Harcourt, Brace & World.

RAJCHMAN, J. [1985]: 'Philosophy in America', in J. Rajchman and C. West (*eds.*), *Post-Analytic Philosophy*. Columbia University Press.

RORTY, R. [1979]: *Philosophy and the Mirror of Nature*. Princeton University Press.

TRIPLETT, T. [1986]: 'Relativism and the Sociology of Mathematics: Remarks on Bloor, Flew, and Frege', *Inquiry*, 29, pp. 439–50.

WACHTERHAUSER, B. [1986]: 'Introduction', to B. Wachterhauser (*ed.*), *Hermeneutics and Modern Philosophy*. State University Press of New York.

Brit. J. Phil. Sci. 39 (1988), 367–373 *Printed in Great Britain*

Popper on Induction

ANDREW J. SWANN

ABSTRACT

The controversy surrounding Popper's proposed solution to the problem of induction is beginning to display many of the symptoms of being interminable. For decades the discussion has continued, apparently without any progress being made. Again and again, Popperians and their critics have accused each other of 'missing the point'.

The essay attempts to explain what exactly *is* 'the point' of the problem of induction, and asks whether Popper does indeed miss it. An answer is proposed, and on this basis an explanation for the puzzling interminability and emptiness of the above dialogue is put forward.

---

1 It is quite easy, and perhaps entertaining, to make an unsympathetic interpretation of Popper's solution to the problem of induction. For example, in his ([1983], p. 28), he defends his solution by saying that non-demonstrability 'never worries the critical rationalist'. This easily lends itself to what may be called the Stoic interpretation; Hume's discovery that all our theories about the world are without foundation is not a problem for Popper, simply because he does not *want* his theories to be justified, and that is an end to it. It was perhaps Epictetus who got the most mileage out of that most useful tautology, 'We can all have everything we want (provided that we want little enough)'. If a man on a sinking ship announces that he has solved the problem of death by joining Exit, it is doubtful whether his fellow passengers would agree, even if they thought their imminent death could not be avoided. I will argue that this analogy is inadequate, because the Stoic interpretation is inadequate.

Again in ([1983], pp. 21–2), 'To have some "foundation", or "justification", may be important for a belief; but it is not the sort of thing we should require for a conjecture or a hypothesis . . .' Let Hume's problem be stated as follows: there are propositions on which we must act, and yet which cannot be justified. Then the above passage, together with Popper's other writings, could be seen by the cynical as 'solving' the problem of induction by calling these propositions 'conjectures' when it is time to justify them, and 'background knowledge' when it is time to act on them (or perhaps 'the light of our critical discussion'); (see [1963], p. 240; [1974], p. 1025–6).

My own interpretation makes use of Professor Susan Haack's suggestion that in Popper's philosophy, much of the work traditionally done by 'beliefs' is done instead by 'held theories'. Popper seems to say that if we are faced with a number of competing theories, we should hold the theory that appears to be the best available (so far, in the light of our critical discussion), and that it is rational to act on that theory. He adds that this is not tautological. (See [1974], pp. 1025–6; [1983], pp. 21–2. To make this interpretation, we must use the word 'hold' as follows: a person S holds a theory p if, and only if, S will act on P rather than its competitors. Notice that the truth of the statement 'S holds p' is problem-situation relative, since p may compete with different rival theories in different action situations.) He later defends this by defending the method of proposing a plurality of guesses and eliminating the worst of them. (See [1973], pp. 365–6.) An ambiguity between the terms 'best/worst' and '*apparently* best/worst' becomes evident, but we will assume that it is purely verbal and unimportant. We can state Popper's theory, then, as follows:

F = 'Given a set of competing theories, we should hold the best theory.'

For some reason, many people seem to think that we should try to avoid acting on unjustified conjectures. Popper offers a rival to this popular notion; there are theories on which we should act, 'best' theories, even though *all* theories about the world are unjustified. (Whether this is merely a verbal disagreement will be considered later.) Hence Hume's problem as we formulated it is not a problem for Popper, so he can hardly be accused of pretending to solve it by calling our theories by different names in different contexts. The Stoic interpretation is also to be rejected. Popper is not in the same situation as the man on the sinking ship, because doing without justification is not as bad as doing without air. There are still propositions on which it is rational to act. However, they are 'best', not 'justified'. If the man on the ship had discovered that human beings do not need air, we would have been nearer the mark. ('I cannot understand how the Aerobist myth has survived for so long!')

Although our earlier formulation of Hume's problem may now be dismissed, Popper still seems vulnerable to Hume's original argument. Whatever F is, it is not a tautology. Naturally we ask how F is to be justified. Popper answers: 'I do not attempt to "justify" the method—I am not a justificationist. On the contrary, I have always stressed that it *may* not succeed; and I shall be surprised but happy if anybody suggests a better or more rational method...' ([1972], pp. 365–6). Popper seems to be saying that we do not need to justify it; we should hold that theory until a better one is found. But that argument *uses* F! The defence is worthless because it presupposes the innocence of the defendant. This was precisely Hume's criticism of any possible defence of induction.

2 Not all problems are questions. Any unacceptable state of affairs represents a problem. The problem of induction, as traditionally experienced, is caused by a rather convincing argument that (roughly) all intellectual endeavour is in vain. It seems to me that Popper's work does represent a step towards solving this problem. Consider the following observation:

> Any justified belief can be deduced from some other justified beliefs, unless it is in the set X (where X contains consistent statements of experience, and necessary truths).

This simple philosophy has been used to combat lies and superstition for centuries, with the merciless question, '*How can you justify that?*' This is the source of all responsible belief; it is therefore the blade with which we cut our speculative theories down to the bloody bone of truth.

The above theory (call it NR) has some plausibility, but it is in fact false. Fortunately, most of the work it does can sometimes be done without it (*e.g.* exposing lies, improving our theories). This theory is very closely modelled on what Popper calls 'uncritical rationalism', and he gives an excellent criticism of it ([1945] p. 230).[1] NR itself cannot be deduced from elements of X; therefore, if an infinite regress is to be avoided, NR is unjustified, or else it is false. NR, then, is self-defeating. (NR stands for Narrow Rationalism; it restricts rationality to some narrow set X.) Even if we include NR *ad hoc* in X, the resultant theory, although not self-defeating, is clearly unacceptable. Moreover, once we have rejected it, we do not have to 'justify' this rejection in any narrow rationalistic sense.

I suggest that the problem of induction is caused by the unconscious use of a version of NR in which the set X contains only logical truths and statements of past experience. For the Humean sceptic will not be satisfied until statements about the future can be deduced from elements of the set X together with some principle I which is also to be deduced from elements of X. This is logically equivalent to deducing statements about the future from the elements of X. It is clear that this person can never be satisfied, and that insofar as his dissatisfaction extends to the rationality of acting on theories about the future, it corresponds to narrow rationalism.

Our argument at the end of the last section, then, being an extension of Hume's argument, is misconceived as it is based on NR. The narrow rationalist's 'defence which does not presuppose the innocence of the defendant' is merely the deduction of the theory from something else; and without NR, we do not always require this.

It must be admitted that Popper does not use the word 'justified' in the way that I have used it. NR should be rewritten:

> Any theory that should be held, can be deduced from other theories that should be held, unless it is in X.

---

[1] The theory that I consider is slightly different, and Popper's criticism has been suitably altered.

Popper would agree that this represents a false theory. However, I think he would say that the similar sentence that we considered earlier was true by definition; true but harmless, because whether a theory is justified has nothing to do with whether it should be held. This may appear perverse. Usage seems to connect 'should be held' with 'justified'. However, Popper seems to have concluded that usage connects altogether too many things that should be kept apart. If we say that any theory that should be held is, by definition, justified; and if we add that justified beliefs can, by definition, be proven; if we accept these definitions, we are already committed to narrow rationalism. (The proof of any theory must obviously presuppose other theories.) Either definition by itself is harmless and without content, but taken together the two definitions lead to NR. The word 'justified' is sometimes also associated with other theories with which Popper disagrees, for example in the field of probability. It appears that false theories may be 'written into the language', and when they are rejected they leave holes that can be patched up in different ways. In the above case, Popper's way is to reject the first definition and retain the second; he concedes the word 'justified' to the narrow rationalist. Only constructions from X can be justified, but theories that should be held are not necessarily limited to those theories which are justified. Using this terminology, NR is the step from the truth of 'p is unjustified' to the falsity of 'p should be held'; presumably this corresponds to the 'justificationist' view to which Popper sometimes refers.

If my interpretation is correct, it seems that Popper is not guilty of linguistic perversity for its own sake when he offers 'best' theories instead of 'justified' ones, but that there is a genuine difference of ideas behind the different words.

The above argument is not to be confused with the so-called 'analytic' solution to the problem of induction, as expounded for example by Paul Edwards in his essay, 'Russell's doubts about induction'. (See Swinburn [1974], pp. 26–47.[2]) Using Russell as a representative of Humean doubt, Edwards does argue that the Humean demand for justification involves a confusion of different senses of the words being used. However, as the name suggests, the analytic argument presents the problem of induction as being somehow empty, a pseudo-problem, nothing more than a sort of intellectual cramp in the mind of Bertrand Russell. I propose on the contrary that Russell's confusion of meanings is synthetic rather than analytic, in the sense that it is associated with the synthetic theory NR; and that Russell did not retain NR *only* because it was written into his terminology. For Russell, NR was an essential ingredient of epistemological responsibility, and he was wary of what he regarded as the slippery slope of deviating from it; 'It is therefore important to discover whether there is any answer to Hume within the framework of a philosophy that is wholly or mainly empirical. If not, there is no intellectual difference between sanity and insanity. The lunatic who believes he is a poached egg is to be condemned solely on the ground that he is in a minority . . .' (Russell [1946], p. 699.) That Russell's empiricism corresponds to NR as we have defined it, is made clear in the following passage from the same essay: '[The principle of induction] must therefore be, or be deduced from, an independent principle not based upon experience. *To this extent*, Hume has proved that *pure empiricism* is not a sufficient basis for science' ([1946], p. 699, my italics).

---

[2] See also two other essays in that volume, Wesley C. Salmon's 'The Concept of Inductive Evidence', and J. O. Urmson's 'Some questions concerning validity', for criticism of Edwards's position. (These arguments are summarized in the editor's preface.)

3 I have interpreted Popper as refuting Narrow Rationalism, the cause of the problem of induction as stated by Hume. This explains some of Popper's otherwise puzzling arguments offered in defence of his solution to that problem. We must now examine the solution itself.

In 'The Logic of Scientific Discovery', Popper argues that Hume's problem of induction is caused by too inflexible an attitude towards empiricism, which leaves science unable to cope with 'Hume's realization of the inadmissibility of inductive arguments' ([1959], p. 42). His controversial solution to Hume's problem is to interpret empiricism—'the thesis that experiment alone can decide upon the truth or falsity of scientific statements'—as being satisfied by falsifiability alone, for 'the method of falsification presupposes no inductive inference, but only the tautological transformations of deductive logic whose validity is not in dispute' ([1959], p. 42).

In accepting Hume's criticism of induction, Popper certainly looks as though he is accepting a narrow rationalist argument. If he is, then his proposed solution to Hume's problem must be rejected. If Humean narrow rationalism is conceded, then *ex hypothesi* Popper's falsificationist empiricism cannot provide the rationality of science, any more than induction or anything else can. This accords well with G. J. Warnock's criticism of Popper's book, in which he argues that *if* there is a problem of induction, *then* Popper's view leads to a similar problem. (Warnock [1960].)

The crucial statement of Popper's position occurs in 'Conjectures and Refutations':

> Hume, I felt, was perfectly right in pointing out that induction cannot be logically justified. He held that there can be no valid logical arguments allowing us to establish 'that those instances, of which we have had no experience, resemble those, of which we have had experience' . . . an attempt to justify the practice of induction by an appeal to experience must lead to an infinite regress. As a result we can say that theories can never be inferred from observation statements, or rationally justified by them.
>
> *I found Hume's refutation of inductive inference clear and conclusive.* (Popper [1963], p. 42, my italics)

Now it could be argued that, for Popper, to say that an inference is 'clearly and concisely refuted' is to say only that it is not a deductive inference. If this were the case, then the quoted passage would not be an example of narrow rationalism, because the 'refutation of inductive inferences' would merely be the observation that inductive inferences are not deductive inferences.

This terminology is misleading in that the inference from A to B can be refuted, even while its validator (*i.e.* the proposition 'If A, then B') remains unrefuted. In its defence, it could be argued that this convention would force us to state all the assumptions we use in arriving rationally at a conclusion. (I do not believe this to be an adequate defence.)

In fact, Popper does not seem to follow this convention, at least in his other writings. In 'Realism and the Aim of Science', he writes:

> In order to discuss the proposal that we should consider universal laws as rules of inference, let us agree to say of statements that they can be true or false, and of inferences . . . that they can be valid or invalid . . . We can then say that in all the usual languages . . . a universal statement like 'all men are bipeds' is factually true if, and only if, the transition from 'x is a man' to 'x is biped' is factually truth preserving (or, as we may say, 'factually valid'). ([1983])

This passage shows a tolerant attitude towards inferences with suppressed premises which seems incompatible with the convention discussed above.

The Humean argument that Popper accepts, then, starts with the premise that a synthetic principle of induction cannot be deduced from experience. From this, it is inferred that the inductive principle is refuted! This is a narrow rationalist argument against induction. Moreover, the inductive principle that he hopes to refute does not entail NR, so he cannot present his argument as a *reductio ad absurdum*. It seems that the solution to Hume's problem that is proposed and defended by the Popperians must be rejected.

4 The controversy surrounding Popper's proposed solution to the problem of induction is beginning to display many of the symptoms of being interminable. For decades the discussion has continued, apparently without any progress being made. Again and again, Popperians and their critics have accused each other of 'missing the point'.

I propose that the point of the problem of induction is that narrow rationalism must be rejected. Popper grasps this point, and indeed most of his philosophy, especially his defence of scientific rationality, is enriched by it. This is the strength of his position. This strength is responsible for the tireless defence of Popperianism that has been maintained for so many years. The weakness of Popper's position lies in the fact that, having refuted NR, he proceeds to use the refuted theory to attack induction. This weakness is responsible for the equally tireless opposition to Popper's views on Hume's problem.

Notice that my positive thesis concerning the Popperian strength is in no way undermined by my analysis of the Popperian weakness; nor would it be undermined by the rejection of Popper's theory F, or of his theories on probability.

7, Greenfield Road
Liverpool 13

REFERENCES

Popper, K. R. [1963]: *Conjuctures and Refutations.* 3rd edition (revised), 1969. Routledge and Kegan Paul.

Popper, K. R. [1959]: *The Logic of Scientific Discovery.* 4th impression, 1965. Hutchinson & Co. Ltd.

Popper, K. R. [1972]: *Objective Knowledge.* 1st edition (revised), 1979. Oxford: Clarendon Press.

Popper, K. R. [1945]: *The Open Society and its Enemies.* 5th edition (revised), 1966. Routledge and Kegan Paul.

Popper, K. R. [1974]: *The Philosophy of Karl Popper.* 1st impression, 1974. Schilpp (*ed.*). The Open Court Publishing Co.

Popper, K. R. [1983]: *Realism and the Aim of Science.* 1st impression, 1983. Hutchinson & Co Ltd.

Russell, B. [1946]: *History of Western Philosophy.* 1st impression, 1946. George Allen & Unwin Ltd.

Swinburne, R. (*ed.*) [1974]: *The Justification of Induction.* 1st impression, 1974. Oxford University Press.

Warnock, G. J. [1960]: Review of 'The Logic of Scientific Discovery'. *Mind*, **69**, pp. 99–101.

Brit. J. Phil. Sci. 39 (1988), 375–378 *Printed in Great Britain*

On Cauchy's Notion of Infinitesimal

NIGEL CUTLAND, CHRISTOPH KESSLER, EKKEHARD KOPP AND DAVID ROSS

The historical literature of nineteenth century mathematical analysis contains much discussion of Cauchy's conception of the continuum, and of the nature of his infinitesimals in particular. In a contribution to this debate, John Cleave [1972], following Lakatos [1978], suggests that Cauchy's 'dynamic' conception of infinitesimals can be accommodated within models of the nonstandard real numbers. Cleave interprets Cauchy's infinitesimals as values of null sequences at infinite integers: an infinitesimal $x \varepsilon {}^*\mathbf{R}$ is **accessible** if there is a **null sequence** (s_n) (i.e. $s_n \to 0$) such that $x = s_M$ for some infinite M. (See Cleave [1972] for the basic definitions).

Cleave states (Theorem 5) that there are non-standard models ${}^*\mathbf{R}$ in which every infinitesimal is accessible; in his proof he claims that this is the case whenever ${}^*\mathbf{R}$ is the ultrapower ${}^{\mathbf{R}^\mathbf{N}}/D$, with D a non-principal ultrafilter on N.

In this note we show that this claim is correct if and only if D is a special kind of ultrafilter—known as a P-point (defined below). Given the continuum hypothesis (CH) or even Martin's Axiom (MA) P-points can be shown to exist; but not all ultrafilters are P-points. Thus, given CH (or MA), Cleave's Theorem is true, but requires a more careful proof. On the other hand, it is consistent with the axioms of set theory (ZFC) that there are no P-points. These matters are discussed in Jech [1978], pp. 257–8. Thus it cannot be shown in ZFC that there are models ${}^{\mathbf{R}^\mathbf{N}}/D$ in which all infinitesimals are accessible.

Definition: An ultrafilter D on N is a **P-point** if whenever

$$N = \bigcup_{n \varepsilon N} X_n$$

is a disjoint union, then there is a set $A \varepsilon D$ such that **either** $A \subseteq X_n$ for some n **or** $A \backslash X_n$ is finite for each n.

It is easy to prove the following:

Lemma: D is a P-point if and only if whenever $x = [(x_n)] \approx O$ there is an $A \varepsilon D$ such that $x_n \to O$ along A.

---

* The authors are grateful to a referee for helpful comments regarding the reference to Cauchy's [1853].

(Here we write $[(x_n)] = \{(y_n)\ \mathbf{R}^{\mathbf{N}}: (x_n) \equiv (y_n)\}$ where $(x_n) \equiv (y_n)$ means that $\{n: x_n = y_n\} \varepsilon D$; and $x \approx O$ means that x is infinitesimal.)

Now we can establish:

Theorem: D is a P-point if and only if every infinitesimal in $\mathbf{R}^{\mathbf{N}}/D$ is accessible.

Proof: First suppose that D is a P-point, and let $x = [(x_n)] \approx O$. Take $A \varepsilon D$ as given by the lemma, so that $x_n \to O$ along A. Define

(s_n) by $$s_n = \begin{cases} x_n \text{ if } n \varepsilon A \\ O \text{ otherwise} \end{cases}.$$

The (s_n) is clearly a null sequence with $x = s_\omega$, where $\omega = [(1,2,3,\ldots)]$, which is plainly infinite.

Conversely, suppose that every infinitesimal in $^{\mathbf{R}^{\mathbf{N}}}/D$ is accessible. Let $x = [(x_n)] \approx O$; we show that $x_n \to O$ along some set $A \varepsilon D$.

Let $y_n = x_n + a_n$, where (a_n) is chosen so that $a_n \to O$ and $y_n \neq y_{n'}$ if $n \neq n'$. Then $y = [(y_n)] \approx O$, so $y = s_M$ for some null sequence (s_n); thus there exists $A \varepsilon D$ with $y_n = s_{m_n}$ for $n \varepsilon A$.

Observe that if $n, n' \varepsilon A$ and $n \neq n'$, then $m_n \neq m_{n'}$, so that $m_n \to \infty$ along A, and consequently $y_n (= s_{m_n}) \to O$ along A. So $x_n \to O$ along A, as required: hence D is a P-point.

The proof contains the following characterisation of P-points:

Corollary: Let $\omega = [(1,2,3,\ldots)]$. Then D is a P-point if and only if every infinitesimal in $^{\mathbf{R}^{\mathbf{N}}}/D$ has the form s_ω for some null sequence (s_n).

Remarks:

(i) The historical interest in the nature of Cauchy's infinitesimals usually centres on the 'mistakes' contained in his famous 1821 treatise **Course d'Analyse** (Cauchy [1821]). His assertion that the limit of a convergent series of continuous functions is again continuous has aroused particular interest, since Cauchy first ignored the counter examples drawn from Fourier series by Abel (1826) and others, and only admitted 'restrictions' to his theorem in Cauchy [1853] after the publication by Stokes (1847) and Seidel (1848) of similar results based on the concept of uniform convergence. Cauchy's paper does not refer to the results of Stokes or Seidel, and his modification of his theorem is anything but clear if we interpret his conception of the continuum as identical with the 'Weierstrassian' concept. Abraham Robinson [1966] first discusses Cauchy's original 'theorem' and its reformulation from the perspective of non-standard analysis, identifying Cauchy's continuum with the nonstandard reals *R. Imre Lakatos [1978] points out that Cauchy's text suggests a more 'dynamic' conception of the continuum: he allows 'moving points' such as $(1/n)$ as infinitesimals, and it is precisely at these points, Cauchy argues, where the 'counter examples' to his theorem fail (see also Spalt [1983]). The purpose of Cleave's [1972] is to elaborate the claim that *R provides a framework in which Cauchy's theorem can be rigorously stated and proved. According to Cauchy's definitions the concept of infinitesimal is subordinate to that of 'variable quantity' (which, as Cleave shows in his [1979], remains a theme of analysis textbooks right up to the 1920's):

> 'On nomme quantité **variable** celle que l'on considéré comme devant recevoir successivement plusieurs valeurs différentes les unes des autres....'
>
> 'Lorsque les valeurs numériques successives d'une même variable décroissent indéfiniment, de manière à s'abaisser au-dessous de tout nombre donné, cette variable devient ce qu'on nomme un **infiniment petit** ou une quantité **infiniment petite**. Une variable de cette espèce a zéro pour limite (Cauchy [1821], p. 4)

These considerations lead Cleave to his definition of 'accessible' infinitesimals and the search for nonstandard models which contain only accessible infinitesimals.

(ii) It is clear that if D is not a P-point then we can find infinitesimals $[s]$ with $s_n \neq 0$ along A for all $A \varepsilon D$. This argues against Cleave's suggestion that Cauchy's notion of infinitesimal could have been made rigorous on the basis of the reduced product $^{\mathbb{R}^{\mathbb{N}}}/F$, where F is the Fréchet filter on N.

(iii) Cleave's analysis of Cauchy's proof is of course unaffected by the above considerations. Taking Cleave's definition of c-convergence as that intended by Cauchy, one is led to conclude that Cauchy's technique of restricting attention to the 'infinitely small neighbourhood' of a point x_0 of c-convergence of the series (which, as Cleave shows, becomes a 'point of uniform convergence' in Weierstrassian terms) allows him to prove a more general result than Seidel's: the two (Weierstrassian) notions coincide when the interval under discussion is compact.

Dept. of Pure Mathematics,
University of Hull

REFERENCES

CAUCHY, A. [1821]: *Cours d'Analyse*, Paris.

CAUCHY, A. [1853]: Note sur les séries convergentes, etc. *Comptes Rendus*, 36, pp. 456–9; *Oeuvres* (Ser. 1) vol. 12, pp. 30–36.

CLEAVE, J. P. [1872]: 'Cauchy, Convergence and Continuity'; *The British Journal for the Philosophy of Science*, 22, pp. 27–37.

CLEAVE, J. P. [1979]: 'The concept of variable in nineteenth century analysis', *The British Journal for the Philosophy of Science*, 30, pp. 266–77.

JECH, T. [1978]: *Set Theory*, Academic Press, N.Y.

LAKATOS, I. [1978]: 'The Significance of Non-Standard Analysis for the History and Philosophy of Mathematics' in: *Imre Lakatos: Philosophical Papers*, Volume 2, C.U.P.

Brit. J. Phil. Sci. 39 (1988), 379–388 *Printed in Great Britain*

DISCUSSION

Was Descartes a Liar? Diagonal Doubt Defended

Arguably, among the most discussed philosophical issues in the history of philosophy have been the Liar paradox and Descartes's *cogito* argument. That these should turn out to be variants of a common underlying puzzle is admittedly a surprising and even eccentric claim at first sight, but it was the burden of my paper 'Descartes's Diagonal Deduction' [1983] to argue at length for just this view (for further details, see Slezak [1985]). If nothing else, the thesis can claim to be a bold conjecture in the Popperian sense of sticking its neck out. To my knowledge, no one had observed this connection before, and this would be surprising if my claim were correct. On its own, of course, such 'boldness' is not necessarily a virtue for it is an attribute shared by all insane and unorthodox theories. Merely remaining unrefuted is no virtue either, for this might happen by default, as it were, just through neglect. Like Norman Bloom's proof of God's existence from the Manhattan telephone directory (Sagan [1979]), one's theory might remain unrefuted simply because no one bothers to take up the challenge. Accordingly, I am grateful to Roy Sorensen [1986] for his recent challenge to my claims concerning Descartes's argument, but I believe these claims survive unscathed, as I should like to show here.

It is difficult to see why Sorensen takes his specific detailed comments to constitute a criticism of my account of the *cogito* in view of the fact that in one place he explicitly endorses my claim concerning the structural parallel between the *cogito* and the Liar ([1986], p. 347), and in another he concedes the point that I cannot coherently doubt 'I doubt this sentence' ([1986], p. 350). I take these very claims to be the central and novel features of my account as a reconstruction of the logical structure of the *cogito*, and it is therefore obscure why Sorensen should simultaneously argue that 'Slezak's *cogito* is not Descartes's *cogito*' ([1986], p. 350). His apparent unwillingness to attribute the analysis to Descartes leaves a prima facie mystery—namely, why such a surprising congruity should hold and yet not be credited to Descartes. In the end, I want to show that Sorensen's inconsistency here derives from a confusion over the status and import of a logical reconstruction such as the one I propose.

Before examining the details of Sorensen's criticisms, it is important to understand that his specific concern with the paradoxical sentences which I claim are implicated in the *cogito* argument may tend to misrepresent the full force and interest of the account I offer. The case in favour of the reconstruction I propose gains strength from convergent arguments of an independent philosophical kind. That is, quite apart from the questions of textual exegesis, Descartes's concerns with thought and the mind permit deploying independent philosophical arguments to illuminate the problem. Furthermore, I take it to be a considerable virtue of the diagonal account of the *cogito* that it is, in fact, consistent with many important insights already available in the vast literature concerning it. My claim has been that the diagonal reconstruction of the *cogito* actually captures these insights by revealing their deeper underlying structure.

Turning to the focus of Sorensen's criticism, it must be noted at the outset that Sorensen does not discuss my arguments entirely in the terms in which I had presented them. Instead he departs from my formulations in significant ways, introducing certain other formulations and digressing into lengthy discussions of such related paradoxical propositions and their apparent implications. Whatever the intrinsic and undoubted interest of these matters, Sorensen does not explicitly defend their relevance to understanding Descartes's argument or my account of it. Clearly, it is not quite fair to make criticisms of my particular account by basing them on quite different formulations which I believe introduce extraneous considerations. In this regard, while Sorensen certainly makes interesting remarks on the problems arising with his own peculiar sentences, it only serves to confuse matters to introduce them in what purports to be a critique of my specific account of Descartes. Thus, Sorensen (pp. 347, 348) discusses at length so-called Buridan sentences such as

(a) It is not the case that I believe (a)

My own account is in terms of a sequence of sentences reflecting the process of doubt, culminating in a sentence of the following form

(x*) I doubt (x*)

Whatever the similarities, the differences in these formulations are important if only because of the concern to capture Descartes's specific reasoning from the specific textual evidence for it. Sorensen's interesting reflections on the implications of (a) and its apparent 'anti-incorrigibility' are of no help in this connection, and nor do they reveal any shortcomings of my account on this score. Among the virtues of my particular reconstruction of the *cogito* argument in terms of the reasoning leading to (x*), I claimed that, purely as a matter of exegesis, it provides a remarkably faithful way of capturing Descartes's literal text. Many of Descartes's remarks which have appeared either anomalous or insignificant, can be seen to have a precise and literal importance on this reconstruction.

A notorious feature of the literature on the *cogito*, particularly in the analytic tradition, has been its extreme uncharitability. Accounts of the *cogito* have been supported at the cost of charging Descartes with a host of sins such as having chosen the verb '*cogitare*' inappropriately, having 'smuggled' more content into his result than is warranted, and not having meant exactly what he said. If this were not unkind enough, the arguments invariably attributed to Descartes are unremarkable and unlikely candidates for what a great philosopher took to be his deepest insight. Now, I have claimed that a further elegant feature of the diagonal reconstruction of the *cogito* is that it permits a literal and charitable reading of the text, which also permits according Descartes a philosophically worthy insight. There is no point rehearsing the substance of these claims, for my concern here is only to note that a sensitivity to the precise textual issues is vital for assessing the value of any putative reconstruction. Through discussing variants or sentences somehow related to my (x*), Sorensen's remarks are of doubtful relevance to my account and in any case miss the full force and significance of the textual evidence I adduce supporting it.

Thus, after lengthy remarks of unclear relevance to my specific concerns, Sorensen reaches the critical conclusion that the anti-incorrigibility of his own sentence (a) makes it an adequate foundation for knowledge ([1986], p. 348). Now, firstly the question of whether or not such a proposition can, in fact, serve as an adequate foundation for knowledge is not the same as the question of whether Descartes thought it was. Secondly, as I have complained, the sentence (a) is, in any case, not one which plays a role in my own reconstruction of the *cogito* argument and, therefore, it shortcomings are not mine.

Somewhat closer to my own (x*), Sorensen considers the sentence

(b) I believe (b) is false

However, Sorensen again departs in significant ways from my own reconstruction of Descartes's reasoning in the process of doubt. Recapitulating briefly, I claimed that my account in terms of (x*) provided little more than a restatement of Descartes's own text leading from doubting the external world to a proposition which attempts to doubt itself and fails, thereby appearing to be incorrigible. Attempting to doubt (x*) makes it true; or alternatively, considering it to be false is to say that it is not the case that I doubt (x*), which is to say that I believe (x*), making it apparently indubitable. By contrast with available accounts, this reconstruction appears to make literal sense of remarks by Descartes which had hitherto been ignored in the literature. Specifically, Descartes explains that he came to know certainty at the same time as he came to doubt and, moreover, the objects of doubt and certainty were different:

> You no sooner showed me the small amount of certainty which we have as to the

> existence of things which are only known to us by the evidence of the senses, than I commenced to doubt of them, and that sufficed to make me know doubt and at the same time my certainty of it, in such a way that I can affirm that as soon as I commented to doubt I commenced to know with certainty. But my doubt and my certainty did not relate to the same object; my doubt regarded things only which existed outside me, my certainty concerned me and my doubt. (Haldane and Ross (*eds.*) [1911], I, p. 325)

On my reconstruction, we can see clearly why the certainty should be simultaneous with the doubt just as Descartes says; it is because the process of doubt itself *produces* the certainty through the peculiarities of the proposition (x*). Notice in this connection how precisely this captures Descartes's words where he says 'and that sufficed to make me know doubt and at the same time my certainty of *it*', where the word 'it' clearly refers to 'doubt'. That is, Descartes's certainty was only in relation to his own doubt itself. It is this peculiar reflexivity which my account explains while giving literal sense to Descartes's remarks. This construal of Descartes's remarks is confirmed explicitly in his following sentence. Here Descartes explicitly draws attention to the different objects of doubt and certainty, his doubt concerning the external world but his certainty concerning *himself and his doubt*. It is a noteworthy fact that other attempts to uncover the logical structure of the *cogito* have had to ignore such explicit clues to Descartes's intentions. Be that as it may, my concern here has been to draw attention to certain features of the diagonal reconstruction which cannot be assessed by considering some other sentences and their implications.

Returning to Sorensen then, we see that it is in relation to *his* own sentence (b), that he derives certain consequences, and on this basis he concludes directly that 'this reading also leaves the *cogito* unserviceable' ([1986], p. 349). Furthermore, he continues

> In addition to rendering the *cogito* unserviceable, Slezak's interpretation would undermine Descartes's theism. (p. 349)

Notice firstly we see here that Sorensen explicitly assimilates his own quite different formulations of the problem to 'Slezak's interpretation'. Secondly, it is not entirely clear what Sorensen intends by the notion 'unserviceable', though it is presumably for the purpose of making an adequate foundation for knowledge. It must be said, however, that the serviceability of a proposition in this regard cannot serve directly as a criterion for judging the exegetical question of whether it captures Descartes meaning, for we have good reasoning nowadays to think that *no* sentence in principle could serve this purpose. Of course, the relevant question is only whether Descartes *thought* some sentence might play such a role. Admittedly, such independent philosophical arguments are by no means always irrelevant to the assessment of textual questions and, indeed, in my own account I have relied heavily on just such considerations as we will see presently. My point here is only that such independent philosophical arguments must be brought to bear judiciously and their relevance explicitly argued. In the present case, I am suggesting that the matter of 'serviceability' cannot help adjudicate the question at issue concerning the logical structure underlying Descartes's reasoning. As we will see, this is because Sorensen's conception of serviceability requires that the analysis in question plays a role in Descartes's *conscious* deliberations, whereas my claim for the diagonal account is only that it has the status of the formal schema *implicit* in the reasoning.

Sorensen's misconceived strategy for attacking my account of the *cogito* is perhaps most clearly seen in his attempt to deploy an argument concerning Descartes's theism. Keeping in mind that Sorensen is still discussing the implications of his own sentence (b), he claims that 'Slezak's interpretation' would undermine Descartes's theism, explaining

> For if Buridan sentences are semantically satisfactory for use in the *cogito*, they can also be used to establish the logical impossibility of God's existence. ([1986], p. 349)

As a matter of methodology, in the absence of any textual or other support, it is surely irrelevant to the understanding of Descartes to adduce properties of an explication which 'can also be used' for some other purposes. However, Sorensen introduces such extraneous considerations and puzzles concerning God and omniscience citing an article by Patrick Grim [1983] which appears to have no bearing whatever on any of Descartes's concerns. Indeed, in the one place that Sorensen explicitly draws a parallel between certain paradoxical sentences concerning God's beliefs and my sentence (x*), he notes that 'the only noteworthy difference' is the substitution of 'God' for 'I'. However, Sorensen has clearly missed the very central thesis of my argument which was to bring into relief the essential role of the first person, introspection and self-reference in Descartes's reasoning. Relevance aside, the substitution of 'God' for 'I' cannot be assumed inessential to the question of understanding Descartes's argument or, more particularly, my account of it. Sorensen's strategy is indicated further in the following remarks

> A second advantage of recasting Grim's argument in Buridanean terms is that Grim's insight can then be easily embedded in the thoughts of a Slezakian Descartes. We can picture Descartes wondering how God would respond to 'I do not believe this sentence'. Since 'I' now refers to God, the proposition expressed will be equivalent to one of Grim's 'Divine Liars'.
>
> Given that Descartes viewed the *cogito* in the way Slezak describes, it is fairly clear as to how he should have resolved his wonderment. ([1986], pp. 349–50)

To imagine Descartes asking such questions about God's beliefs and speculating further about how he might have answered them is to introduce irrelevant issues together with gratuitous speculations about fictional problems. Such considerations cannot be appropriately deployed to assess the merits of a rational reconstruction of Descartes's reasoning. Sorensen's reference to Descartes's 'wonderment' above makes it appear as if genuine textual questions were being raised, but it must be clearly emphasized that this is a pure fantasy constructed around Sorensen's independent concerns which have neither textual nor direct philosophical support. Sorensen's strategy here is to consider how Descartes might have reacted, given what he (Sorensen) takes to be the implications of my account; he concludes that, given my account, Descartes would have had grounds to establish the logical impossibility of God's existence ([1986], p. 349); hence, my account is implausible as an account of Descartes's intentions.

A clue to Sorensen's mistake is revealed in his suggestion that the Buridanean arguments raised by Grim can be 'easily embedded in the thoughts of a Slezakian Descartes' ([1986], p. 349). The very idea of a 'Slezakian Descartes' in Sorensen's sense and the question of how such a Descartes might have reasoned can only make sense on the assumption that such a Descartes would have had access to the explication being proposed. Evidently, it is only on such a tacit assumption that it could make sense to speculate about what Descartes might have thought about God, given my account. That is, Sorensen's story rests on Descartes being in a position to perceive the alleged implications of my account for his theism. Why else picture Descartes wondering anachronistically about Grim's paradoxes concerning God's omniscience? By contrast, my account purports only to reveal features of Descartes's argument which are *implicit*, in essentially the way that the rules of grammar are implicit in an utterance. On the other hand, Sorensen speculates about what Descartes might have actually thought or believed given certain views, which is to assume these views to be available to him in a significantly different sense. Clearly, my diagonal account of the *cogito* can have no implications whatever for any of Descartes's beliefs as Sorensen suggests. The point can be easily missed in view of the significant ambiguity of the idea that 'Descartes viewed the *cogito* in the way Slezak describes'. The ambiguity corresponds essentially to the transparent and opaque readings of the intentional context. Descartes could simultaneously have viewed the *cogito* in the way I describe it in one sense, and yet not have viewed it in that way in the other sense.

It is on the basis of these confusions that Sorensen suggests 'Slezak's interpretation of Descartes's *cogito* leaves Descartes vulnerable to embarrassing objections' ([1986], p. 350). Triumphantly, he observes further that 'Descartes displays little interest in answering these questions' ([1986], p. 350). However, there is an extreme irony in this latter remark which Sorensen takes to be a criticism of my account. The obvious reason Descartes shows no interest in answering Sorensen's questions is that they are unlikely to have occurred to him since they have nothing whatever to do with his actual concerns. Without impugning the philosophical interest of the issues Sorensen raises, I have tried to indicate the importance of distinguishing the quite separate enterprises which may be involved here. On the one hand there is the issue of Descartes's meaning and intentions and on the other hand there is the issue of the independent philosophical puzzles surrounding Descartes's concerns. Admittedly, as a matter of methodology, the two cannot be entirely separated, as indeed I have argued at length in my paper, for the independent philosophical considerations can serve in a heuristic capacity as clues to Descartes's meaning. However the use of such independent philosophical considerations must be cautious and conform to certain criteria of adequacy. Clearly, not all the questions raised in a philosophical domain will be equally relevant in reconstructing the thought of some philosopher. In the present case, there is not the slightest argument offered for the bearing of Sorensen's problems on the historical Descartes and the textual or exegetical evidence.

In observing that Descartes shows no interest in answering certain questions, Sorensen even appears to be taking the *absence* of textual evidence to be *support* for his view. Obviously there are an infinite number of such questions which Descartes shows no interest in answering and his silence on these is no evidence for anything. Sorensen writes

> Rather than concluding that Descartes overlooked these points, it is more likely that the objections are irrelevant to the *cogito* that Descartes had in mind. Since the objections are relevant to the *cogito* Slezak had in mind, I think we should conclude that Slezak's *cogito* is not Descartes's *cogito* ([1986], p. 350)

Sorensen's confidence that Descartes should have seen these implications (if my account is correct) is further confirmation of the fundamental confusion I have noted above. Here we see Sorensen speaks of what Descartes might have overlooked and 'the *cogito* that Descartes had in mind' juxtaposed with 'the *cogito* that Slezak had in mind' ([1986], p. 350); elsewhere he is concerned with whether 'Descartes viewed the *cogito* in the way Slezak describes' ([1986], p. 350), and says that 'a Slezakian Descartes would not be in a position to shelve the atheistic argument' ([1986], p. 350). This way of talking makes it clear that Sorensen is relying on attributing to Descartes just the kind of conscious belief concerning my account which must be anachronistic and unwarranted. The only sense in which one would want to say that this analysis of the *cogito* is the one Descartes 'had in mind' is a sense in which it reveals the underlying logic by employing concepts which need not themselves even have been available to Descartes. Likewise, Epimenides made a diagonalization whether he knew it or not.

Yet further support for my diagnosis of Sorensen's mistake is seen in his remark that 'Slezak's interpretation of Descartes's *cogito* leaves Descartes vulnerable to embarrassing objections' ([1986], p. 350). It is hard to see how any analysis as such could leave *Descartes* open to objections, unless it is assumed that Descartes himself possessed the analysis in question. To be sure, Descartes not only uttered the argument but was also concerned with its analysis, but could not be thought to have *entertained* my account as Sorensen's argument seems to require.

Again, we see Sorensen write

> So if forced to choose between this version of the *cogito* and theism, Descartes would have lost little by choosing theism. ([1986], p. 350)

Though the issues of theism are derived from the alleged implications of my analysis, Descartes could hardly have been faced with this choice as Sorensen suggests. Again we see that Sorensen clearly assumes Descartes might have actually entertained the diagonal account and its implications, as he did entertain, for example, the syllogistic interpretation. To the extent that Sorensen's entire criticism is founded on this view of Descartes's available choices concerning the proper analysis of the *cogito* argument, he has missed the import and interest of the analysis I have proposed.

Many of Sorensen's locutions strongly suggest a reading which construes the paradoxical sentences as part of the very content of Descartes's thought. Thus he says that my interpretation assigns the paradoxical sentences 'a legitimate role in Descartes's philosophy' ([1986], p. 346). Admittedly this is ambiguous, but it would be a slightly odd way to speak of an account which has the status I have attributed to the diagonal schema. The appropriate status to accord the formal diagonal model is familiar from Chomsky's writings on the methodology of linguistics where he has written of the grammar of a language as the rules and representations constituting the *tacit knowledge* of the idealized speaker-hearer. Observational and descriptive adequacy of a grammar or competence model are achieved if they capture the intuitions of the native speaker. The intuitions are consciously available to introspection but the formalisms of the grammar themselves are not. These latter are taken to be abstract representations realized in unknown 'performance' mechanisms in the brain. It was precisely in this sense that I suggested that the formal self-referential schemata may provide abstract models of deep shared intuitions about ourselves ([1983], p. 35) and the mathematical notions would have a status closely analogous to the formalisms of Chomskyan competence models of tacit knowledge ([1983], p. 21). It is perhaps worth noting here that the appropriateness of this methodological apparatus for appreciating Descartes's insight is borne out by the further textual and philosophical coherence it provides: It is a striking feature of this analysis that it provides a natural resolution to the difficulties regarding Descartes's apparently conflicting remarks concerning the role of intuition and inference in relation to the *cogito*. While speaking of his insight as a most certain inference, in some places Descartes appears to deny its inferential character, describing it more like a sudden vision. The apparent contradiction seen here by commentators is neatly resolved with the methodological distinctions I have indicated, since we can see the precise legitimacy of these alternative ways of describing his insight.

On this understanding of my analysis, it is difficult to see how any of Sorensen's criticisms could be sustained. My defence of the diagonal analysis has been directed at the specific criticisms raised by Sorensen and it is, therefore, both unnecessary and inappropriate to rehearse the full argument here. Nevertheless, it is important to draw attention to one feature of my original account which is disguised by the very focussed nature of the foregoing comments. Quite aside from the specific substance of Sorensen's criticisms, his approach to the *cogito* is characteristic of the entire tradition of philosophical discussion, especially among analytic philosophers, in being focussed directly and narrowly on the *cogito* dictum itself in abstraction from the broader context of reasoning in which it occurs. From Descartes's time onwards, the attempts to uncover the logical structure of the argument have failed to accommodate the broader context of the argument as a whole and I claim it to be a considerable virtue of my account that it naturally permits demonstrating the precise logical significance of these more general features of Descartes's reasoning (see Slezak [1985]). In this respect too I have claimed that we see an elegant convergence of philosophical and textual considerations; what appear to be otherwise insignificant or anomalous remarks by Descartes take on a strikingly literal meaning under the proposed interpretation. Thus, for example, Descartes warns against just such narrowing of focus as has been characteristic of the subsequent discussions saying.

> But for those who, without caring to comprehend the order and connections of my reasonings, form their criticisms on detached portions arbitrarily selected, as is the custom with many, these, I say, will not obtain much profit from reading this Treatise. And although they perhaps in several parts find occasion of cavilling, they can for all their pains make no objection which is urgent or derserving of reply. (HR 1, p. 139)

More specifically, Descartes insists elsewhere on the crucial role of the entire doubting process in the insight:

> My statement that the entire testimony of the senses must be considered to be uncertain, nay, even false, is quite serious and so necessary to the comprehension of my meditations, that he who will not or cannot admit that, is unfit to urge any objection to them that merits a reply. (HR 2, p. 206)

My diagonal reconstruction is based on an enumeration of empirical propositions and in this way actually incorporates this process of denying the testimony of senses into the very logic of the argument. At the same time, this enumeration which is the basis of the diagonalization remarkably conforms to Descartes recommendation that 'order is what is needed: all the thoughts that can come into the human mind must be arranged in an order like the natural order of the numbers' (Kenny (*ed.*) [1970], p. 5).

In general, the neglect of the broader structure of Descartes's reasoning has undoubtedly obscured the significance of other textual clues such as Descartes's insistence that the *cogito* is an example of the ancient method of analysis. The method of analysis and synthesis to which Descartes alludes typically involves longer chains of reasoning and its relevance has been entirely obscure so long as the logic of the argument was sought narrowly within the confines of the *cogito* dictum. On the other hand, I have tried to show in detail ([1985], p. 208 ff.) how the diagonal account permits taking Descartes at his word and making sense of the argument as an application of the method of analysis.

My concern here is not to defend these specific claims, but only to indicate the importance of broader considerations in assessing the merits of my account of the *cogito*. In the same way I would draw attention to the entirely independent philosophical considerations concerning the self and introspection which I have adduced as providing converging support for a diagonal analysis (see [1983] and [1985]). Whatever the merits of the individual arguments, the very consilience of diverse textual and philosophical considerations provides added support for such each one individually. Undoubtedly, showing the affinity of the *cogito* to the Liar does not resolve all its problems, but at least being able to assimilate it to the familiar paradox helps to explain why the *cogito* should have appeared so perplexing by reducing two puzzles to one.

PETER SLEZAK
University of New South Wales

REFERENCES

GRIM, PATRICK [1983]: 'Some Neglected Problems of Omniscience', *American Philosophical Quarterly*, 20, pp. 265–76.

HALDANE, E. and ROSS, G. (*eds.*) [1911]: *The Philosophical Works of Descartes*. Vols. I and II. Cambridge; Cambridge University Press; reprinted 1973.

KENNY, A. (*ed.*) [1970]: *Descartes: Philosophical Letters*. Oxford: Clarendon Press.

SAGAN, CARL [1979]: 'Norman Bloom, Messenger of God' in *Broca's Brain*. London: Hodder and Stoughton.

SLEZAK, PETER [1983]: 'Descartes's Diagonal Deduction', *British Journal for the Philosophy of Science*, 34, pp. 13–36.

SLEZAK, PETER [1985]: *The Self in Thought and Action*. Doctoral Dissertation, Columbia University. Ann Arbor: University Microfilms International.

SORENSEN, ROY [1986]: 'Was Descartes's *Cogito* a Diagonal Deduction?' *British Journal for the Philosophy of Science*, 37, pp. 346–51.

Brit. J. Phil. Sci. 38 (1988), 389–395 *Printed in Great Britain*

DISCUSSION

Barrow and Tipler on the Anthropic Principle *vs.* Divine Design

In their massive study *The Anthropic Cosmological Principle,* [1986][1] John Barrow and Frank Tipler provide the most comprehensive analysis to date of the so-called Anthropic Principle and its relation to the classic teleological argument for a Divine Designer of the cosmos. According to their analysis, the Anthropic Principle evolved out of the traditional design argument for God's existence, particularly one version of that argument, the eutaxiological version, which was based on the presence of discernable order and mutual harmony in nature in abstraction from any anthropocentric purpose being in view. Although Barrow and Tipler believe that the Darwinian theory of evolution undermined biological, anthropocentric versions of the teleological argument, they contend that contemporary science has only served to accentuate the delicate balance, perceived in the eutaxiological version of that argument, of hightly improbable necessary conditions for the evolution and sustenance of intelligent life which obtain in the universe, and the bulk of their book is devoted to surveying the fields of physics and astrophysics, classical cosmology, quantum mechanics, and biochemistry to illustrate their point. These supply the evidence for what F. R. Tennant [1930], who coined the term *anthropic*, called 'wider teleology'.

Not that Barrow and Tipler are endorsing a design argument; on the contrary, although scientists hostile to teleology are apt to interpret their work as sympathetic to theism and although I have already seen this book cited by two prominent philosophers of religion in support of the teleological argument, the thrust of the book's argument is in the end anti-theistic. As Barrow and Tipler employ it, the Anthropic Principle is essentially an attempt to complete the job, begun by Darwinian evolution, of dismantling the teleological argument by showing that the appearance of design in the physical and cosmological quantities of the universe is just that: an *appearance* due to the self-selection factor imposed on our observations by our own existence. If Barrow and Tipler are correct, then the wider teleological argument of Tennant proves no more effective than the narrow teleological argument of his predecessors.

[1] For a more wide-ranging review of this book see Craig [forthcoming].

That brings us to a consideration of the Anthropic Principle itself. Barrow and Tipler distinguish several versions of the Principle, the most basic and least disputable being the Weak Anthropic Principle (WAP):

> WAP: The observed values of all physical and cosmological quantities are not equally probable, but they take on values restricted by the requirement that there exist sites where carbon-based life can evolve and by the requirement that the Universe be old enough for it to have already done so. (p. 15)

Barrow and Tipler regard WAP as 'in no way speculative or controversial' (p. 16), since it is 'just a restatement . . . of one of the most important and well-established principles of science: that it is essential to take into account the limitations of one's measuring apparatus when interpreting one's observations' (p. 23). For example, if we were calculating the fraction of galaxies that lie within certain ranges of brightness, our observations would be biased toward the brighter ones, since we cannot see the dim ones so easily. Or again, a ratcatcher may say that all rats are bigger than six inches because that is the size of his traps. Similarly, any observed properties of the universe which may initially appear astonishingly improbable can only be seen in their true perspective after we have accounted for the fact that certain properties could not be observed by us, were they to obtain, because we can only observe those compatible with our own existence. 'The basic features of the Universe, including such properties as its shape, size, age, and laws of change must be *observed* to be of a type that allows the evolution of observers, for if intelligent life did not evolve in an otherwise possible universe, it is obvious that no one would be asking the reason for the observed shape, size, age, and so forth of the universe' (pp. 1–2). Thus, our own existence acts as a selection effect in assessing the various properties of the universe. For example, a life form which evolved on an earthlike planet 'must necessarily see the Universe to be at least several billion years old and . . . several billion light years across,' for this is the time necessary for production of the elements essential to life and so forth (p. 3).

Now, we might ask, why is the '*observed*' in the quotation in the above paragraph italicized? Why not omit the word altogether? The answer is that the resulting statement:

1. The basic features of the universe must be of a type that allows the evolution of observers

is undoubtedly false; for it is not logically or nomologically necessary that the universe embrace intelligent life. Rather what seems to be necessarily true is

2. If the universe is observed by observers which have evolved within it, then its basic features must be of a type that allows the evolution of observers within it.

But (2) seems quite trivial; it does nothing to explain why the universe in fact has the basic features it does.

But Barrow and Tipler contend that while (2) appears to be true, but trivial, it has 'far-reaching implications' (p. 2). For the implication of WAP, which they seem to interpret along the lines of (2), is that no explanation of the basic features of the universe need be sought. This contention seems to be intimately connected with what is appropriate to be *surprised at*. The implication of WAP is that we ought not to be surprised at observing the universe to be as it is, for if it were not as it is, we could not observe it. For example, 'No one should be surprised to find the Universe to be as large as it is' (p. 18). '. . . on Anthropic grounds, we should expect to observe a world possessing precisely three spatial dimensions' (p. 247). Or again,

> We should emphasize once again that the enormous improbability of the evolution of intelligent life in general and *Homo sapiens* in particular does *not* mean we should be amazed we exist at all. This would make as much sense as Elizabeth II being amazed she is Queen of England. Even though the probability of a given Briton being monarch is about 10^{-8}, *someone* must be. Only if there *is* a monarch is it possible for the monarch to calculate the improbability of her particular existence. Similarly, only if an intelligent species does evolve is it possible for its members to ask how probable it is for an intelligent species to evolve. Both are examples of WAP self-selection in action.[110]
>
> [110] F. B. Salisbury, *Nature* 224, p. 342 (1969), argued that the enormous improbability of a given gene, which we computed in the text, means that a gene is too unique to come into being by natural selection acting on chance mutations. WAP self-selection refutes this argument, as R. F. Doolittle in *Scientists confront creationism*, L. R. Godfrey (Norton, NY, 1983) has also pointed out (pp. 566, 575).

Here we have a far-reaching implication that goes considerably beyond the apparently trivial WAP. Accordingly, although Barrow and Tipler conflate WAP and the implications thought to follow from it, I want to distinguish these sharply and shall refer to these broader implications as the Anthropic Philosophy. It is this philosophical viewpoint, rather than WAP itself, that, I believe, despite initial impressions, stands opposed to the teleological argument and constitutes scientific naturalism's most recent answer to that argument. According to the Anthropic Philosophy, an attitude of surprise at the delicately balanced features of the universe essential to life is inappropriate; we should expect the universe to look this way. While this does not explain the origin of those features, it shows that no explanation is necesary. Hence, to posit a divine Designer is gratuitous.

Now it needs to be emphasized that what the Anthropic Philosophy does *not* hold, despite the sloppy statements on this head often made by scientists, is that our existence as observers explains the basic features of the Universe. The answer to the question 'Why is the universe isotropic?' given by Collins and Hawking, '. . . the isotrophy of the Universe is a consequence of our existence'

(Collins and Hawking [1973], p. 317) is simply irresponsible and brings the Anthropic Philosophy into undeserved disrepute, for literally taken, such an answer would require some form of backward causation whereby the conditions of the early universe were brought about by us acting as efficient causes merely by our observing the heavens. But WAP neither asserts nor implies this; rather WAP holds that we must *observe* the universe to possess certain features (not that the universe must possess certain features) and the Anthropic Philosophy says that therefore these features ought not to surprise us or cry out for explanation. The self-selection effect affects our observations, not the basic features of the universe itself. If the Anthropic Philosophy held that the basic features of the universe were themselves brought about by our observations, then it could be rightly dismissed as fanciful. But the Anthropic Philosophy is much more subtle: it does not try to explain *why* the universe has the basic features it does, but contends that no explanation is *needed*, since we should not be surprised at observing what we do, our observations of those basic features being restricted by our own existence as observers.

But does the Anthropic Philosophy follow from the Anthropic Principle, as Barrow and Tipler claim? Let us concede that it follows from WAP that

3. We should not be surprised that we do not observe features of the universe which are incompatible with our own existence.

For if the features of the universe were incompatible with our existence, we should not be here to notice it. Hence, it is not surprising that we do not observe such features. But it follows neither from WAP nor (3) that

4. We should not be surprised that we do observe features of the universe which are compatible with out existence.

For although the object of surprise in (4) might at first blush appear to be simply the contrapositive of the object of surprise in (3), this is mistaken. This can be clearly seen by means of an illustration (borrowed from John Leslie): suppose you are dragged before a firing squad of 100 trained marksmen, all of them with rifles aimed at your heart, to be executed. The command is given; you hear the deafening sound of the guns. *And you observe that you are still alive*, that all of the 100 marksmen missed! Now while it is true that

5. You should not be surprised that you do not observe that you are dead,

nonetheless it is equally true that

6. You should be surprised that you do observe that you are alive.

Since the firing squad's missing you altogether is extremely improbable, the surprise expressed in (6) is wholly appropriate, though you are not surprised that you do not observe that you are dead, since if you were dead you could not observe it. Similarly, while we should not be surprised that we do not observe features of the universe which are incompatible with our existence, it is nevertheless true that

7. We should be surprised that we do observe features of the universe which are compatible with our existence,

in view of the enormous improbability, demonstrated repeatedly by Barrow and Tipler, that the universe should possess such features.

The reason the falsity of (7) does not follow from (3) is that subimplication fails for first order predicate calculus. For (3) may be schematized as

$$3'. \sim S\colon (x)\,([Fx \cdot \sim Cx] \supset \sim Ox)$$

where S: is an operator expressing 'we should be surprised that' and F is 'is a feature of the universe', C is 'is compatible with our existence', and O is 'is observed by us'. And (7) may be schematized as

$$7'.\ S\colon (\exists x)\,([Fx \cdot Cx] \cdot Ox)$$

It is clear that the object of surprise in (7′) is not equivalent to the object of surprise in (3′); therefore the truth of (3′) does not entail the negation of (7′).[2]

Therefore, the attempt of the Anthropic Philosophy to stave off our surprise at the basic features of the universe fails. It does not after all follow from WAP that our surprise at the basic features of universe is unwarranted or inappropriate and that they do not therefore cry out for explanation. But which features of the universe should thus surprise us?—those which are necessary conditions of our existence and which seem extremely improbable or whose coincidence seems extremely improbable. Thus, we should amend (7) to read

7*. We should be surprised that we do observe basic features of the universe which individually or collectively are excessively improbable and are necessary conditions of our own existence.

Against (7*), both the WAP and the Anthropic Philosophy are impotent. But which features are these specifically? Read Barrow and Tipler's book. Once this central fallacy is removed, their volume becomes for the design argument in the twentieth century what Paley's *Natural Theology* was in the nineteenth, *viz.*, a compendium of the data of contemporary science which point to a design in nature inexplicable in natural terms and therefore pointing to the Divine Designer.[3]

---

[2] Similarly, the falsity of (6) does not follow from the truth of (5), for (5) may be schematized as $\sim S\colon \sim(\exists x)\,([Mx \cdot \sim Ax] \cdot Ox)$, where M is 'is me', Ox is 'is observed by me' and A is 'is alive'. From this it does not follow that $\sim S\colon (\exists x)\,([Mx \cdot Ax] \cdot Ox)$, which is the negation of (6).

[3] Once the central fallacy is thus removed, Barrow and Tipler's argument in the lengthy quotation in the text seems to amount to little more than the old objection that *any* state of affairs is highly improbable and therefore the obtaining of the actual state of affairs requires no special explanation. But this objection is surely misconceived. What unprejudiced and right-minded

Now Barrow and Tipler will no doubt contend that I have missed the whole point of WAP. For (7*) is true only if the basic features of our observable universe are co-extensive with the basic features of the Universe as a whole. And it may well be the case that the Universe at large does not in fact display the apparent features of design which our segment does. Barrow and Tipler endorse the Many-Worlds Interpretation of quantum physics, but one could also appeal to inflationary models or oscillating models of the Universe in order to generate multiple worlds. If such a wider Universe exists, then it might be argued that all possible universes are actualized and that WAP reveals why surprise at our being in a universe with basic features essential to life is not appropriate.

Objections can be raised against each of the theories proposed for generating many worlds; but even if we conceded that a multiple universe scenario is unobjectionable, would such a move succeed in rescuing as from teleology and a cosmic Designer? This is not at all obvious. The fundamental assumption behind the Anthropic philosopher's reasoning in this regard seems to be something along the lines of

8. If the Universe contains an exhaustively random and infinite number of universes, then anything that can occur with non-vanishing probability will occur somewhere.

But why should we think that the number of universes is actually infinite? This is by no means inevitable, not to mention the paradoxical nature of the existence of an actually infinite number of things. And why should we think that the multiple universes are exhaustively random? Again, this is not a necessary condition of many-worlds hypotheses. In order to elude the teleological argument, we are being asked to assume much more than the mere existence of multiple universes.

In any case, the move on the part of Anthropic philosophers to posit many worlds, even if viable, represents a significant concession because it implies that the popular use of the WAP to refute teleology in a Universe whose properties are coextensive with the basic features of our universe is fallacious. In order to stave off the conclusion of a Designer, the Anthropic philosopher must take the metaphysically speculative step of embracing a special kind of

---

person could possibly regard a chimpanzee's haphazardly typing out the complete plays and sonnets of Shakespeare as equally probable with any chaotic series of letters? The objection fails to reckon with the difference between randomness, order, and complexity. On the first level of randomness, there is a non-denumerably infinite number of chaotic sequences, *e.g.*, 'adfzwj', each of which is equally improbable and which collectively could serve to exhaust all sequences typed by the ape. But the meta-level of ordered letters, *e.g.*, 'crystalcrystalcrystal', need never be produced by his random efforts, were he to type for eternity. Even more improbable is the meta-meta-level of complexity, in which information is supplied, *e.g.*, 'To be or not to be, that is the question.' Hence, it is fallacious to assert that since *some* set of conditions must obtain in the universe, the actual set is in no way improbable or in need of explanation.

multiple universe scenario. That will hardly commend itself to some as any less objectionable than theism.

We appear then to be confronted with two alternatives: posit either a cosmic Designer or an exhaustively random, infinite number of other worlds. Faced with these options, is not theism just as rational a choice as multiple worlds?

Barrow and Tipler demur, maintaining that 'careful thinkers' would not today 'jump so readily' to a Designer, for (i) the modern viewpoint stresses time's role in nature; but since an unfinished watch does not work, arguments based on omnipresent harmony have been abandoned for arguments based on co-present coincidences; and (ii) scientific models aim to be realistic, but are in fact only approximations of reality; so we hesitate to draw far-reaching conclusions about the nature of ultimate reality from models that are at some level inaccurate (p.30). But Barrow and Tipler seem unduly diffident here. A careful thinker will not readily jump to any conclusions, but why may he not infer a Divine Designer after a careful consideration of the evidence? Point (i) is misleading, since the operations of nature always work; at an earlier time nature is not like an *unfinished* watch, rather it is just a *less complex* watch. In any case, the most powerful design argument will appeal to both present adaptedness and co-present coincidences. Point (ii) loses much of its force in light of two considerations: (a) this is a condition that affects virtually all our knowledge, which is to say that it affects none of it in particular, so that our only recourse is simply to draw conclusions based on what we determine most accurately to reflect reality; fortunately, the evidence at issue here is rather concrete and so possesses a high degree of objectivity; (b) Barrow and Tipler do not feel compelled to exercise such restraint when proposing metaphysically speculative but naturalistic accounts of the universe's basic features, *e.g.*, their defense of the 'many worlds' interpretation of quantum physics or scenarios for the origin of the universe *ex nihilo*, which leads one to suspect that a double standard is being employed here.

Hence, the Anthropic Principle notwithstanding, I see no reason why a careful thinker may not, on the basis of the teleological argument, rationally infer the existence of a supernatural intelligence which designed the universe.

WILLIAM LANE CRAIG
Université Catholique de Louvain,
Belgium

REFERENCES

BARROW, JOHN and TIPLER, FRANK (1986): *The Anthropic Cosmological Principle.* Clarendon Press.

COLLINS, C. B. and HAWKING, S. W. (1973): 'Why is the Universe Isotropic?' *Astrophysical Journal* 180, 317–34.

CRAIG, WILLIAM LANE (forthcoming): 'Critical review of *The Anthropic Cosmological Principle. International Philosophical Quarterly.*

TENNANT, R. (1930): *Philosophical Theology.* 2 vols. Cambridge University Press.

Brit. J. Phil. Sci. 39 (1988), 397–401 *Printed in Great Britain*

DISCUSSION

A Note On Schrödinger's Cat and the Unexpected Hanging Paradox

I INTRODUCTION

In Loeser [1984], the unexpected hanging problem is used via analogy to try to clarify possible resolutions of the Schrödinger cat paradox. This stimulated me to consider the reverse—whether the Schrödinger cat paradox (or, more generally, the measurement problem) can shed any light on the unexpected hanging problem. Two analogies presented themselves. Although one can question how strong each is, some light is shed on the unexpected hanging problem and, further, an interesting common theme will be seen to emerge. The unexpected hanging and measurement problems both have unsatisfying or puzzling consequences following apparently sound logic. The first analogy reflects facets of this. The second analogy shows how the introduction of elements of uncertainty can point the way out of each problem. In both cases, the uncertainty is associated with problematical transitions, basic to both problems.

First, we show that there is an analogy to the *process of analyzing* the unexpected hanging problem. This leads to exposure of a problem in some proposed resolutions of the apparent paradox. Secondly, we show there is some parallel between a recent view of the unexpected hanging problem and an explanation of the measurement problem.

We won't take the space here to repeat the unexpected hanging problem statement since it is stated in so many places including Loeser [1984]. There is a large literature on the problem—an excellent review with 40 references is given in Margalit and Bar-Hillel [1983] and there is still a continuing flow of papers. Briefly, the problem is based on ruling out a last day for the occurrence of an announced unexpected event and a backwards-in-days reasoning by the prisoner that eliminates all preceding days. This reasoning is contradicted by the actual occurrence of the unexpected event. The challenge is to find a flaw in what appears to be sound logic on the part of the prisoner. As is well known, Schödinger's cat is associated with the difficulty of understanding the 'wave function collapse' and the role of the registration or measuring system or, possibly, of a conscious observer (after Von Neumann, London, Bauer, Wigner; see d'Espagnat [1976] or Primas [1983] for discussion and references).

2 ANALYSING THE UNEXPECTED HANGING PROBLEM

In quantum physics, the relationship of the measuring system to what is being measured is crucial. This has sort of an analogy to the analyse[1] of the unexpected hanging problem and the problem itself. To see this, we examine a proposed resolution of the unexpected hanging paradox which has several variants of it in the literature on the problem.

The proposed resolution in question (call it R) uses the following argument: after the prisoner concludes reasoning, he/she becomes vulnerable to being surprised because of the very fact of no longer expecting to be hanged. So due to the prisoner changing expectation as a result of reasoning about the situation without fully realizing the consequences, the announcement can be carried out and a difficulty in the prisoner's reasoning is pointed out. This argument holds even, or especially, on the last day.

In the above reasoning of R there is a question as to the latitude in reasoning allowed the prisoner. Whatever we can deduce about the problem should also be available to the prisoner. Thus, the prisoner should be allowed the possibility of reasoning that not expecting to be hanged leads to vulnerability to being surprised. This subverts that lack of expectation, undermining R.

The above shows that R is based on the independence of the reasoning of the analyst and the prisoner (or, more accurately, a privileged status is given to the analyst). A somewhat analogous error in physics would be to ignore the interaction between measuring system and system being measured.

To be more explicit about the interactions between analyst and prisoner and between measuring system and system being measured, we represent each as a sequence. Denote the analyst and prisoner by A and P, resp., and i as a stage in the reasoning process of A and P:

A_1. As a result of the backwards-in-days reasoning there cannot be an unexpected hanging.
P_1. A_1 is available to P$\Rightarrow$P does not expected to be hanged.
A_2. P is vulnerable to being surprised.
P_2. A_2 is available to P$\Rightarrow$P realizes that he/she can be surprised as a result of not expecting a hanging and will drop that lack of expectation.

R carried the reasoning through A_2 and we carried the reasoning through P_2. The question arises as to whether this line of reasoning leads to a finite procedure which resolves the paradox. In fact, resolutions have been proposed around stage P_2—see *e.g.*, the fourth paragraph of Section IV of Loeser [1984]. Examination of some other discussions of this problem has the analyst in an admittedly nonterminating or circular line of reasoning which highlights the apparent paradoxical nature of the problem. This nonterminating reasoning sometimes takes the form of the analyst's reasoning iterating with the prisoner's.

---

[1] By analyst, we mean someone who tries to resolve the problem, *e.g.*, an author of a paper on it.

It is also the lack of a finite procedure in a (simplified) view of the measurement process which leads to the suggestion of a conscious observer for termination. That is, by considering the sequence:

S_1: system
S_2: S_1+measurement system of S_1
S_3: S_2+measurement system of S_2
.
.
.

there appears to be no satisfactory mechanism (via Schrödinger equation, linear wave mechanics) to terminate the sequence with a wave function collapse. Of course, the conscious observer has also met with objections, part of the difficulty of the measurement problem and Schrödinger's cat.

In the measurement problem it is the likeness of the system and the measurement system (both subject to linear wave mechanics) that makes the problem so difficult. The conscious observer was proposed as a solution because of its purported difference (not subject to linear wave mechanics) from the system to be measured. In the unexpected hanging problem, it is the likeness of the reasoning ability of the analyst and the prisoner which causes difficulty with R.

Although it was not necessary to use analogy to find a defect in R, thinking about these parallels did, in fact, stimulate me to think further about R.

3 UNCERTAINTY AS AN ANTIDOTE TO PUZZLING LOGICAL CONSEQUENCES

In the previous section, we faulted R for giving privileged reasoning status to the analyst over the prisoner. In the formulation of Holtzman [1987], the prisoner being able to deduce whatever anybody else can leads to an undecidable aspect. The undecidable aspect represents a flaw in the prisoner's backwards-in-days reasoning and preserves the uncertainty which allows performance of the announcement.[2] (This could be viewed as a fulfillment of an expectation expressed by Meltzer in Meltzer and Good [1965] that a decidability-related notion should play a role in this problem.) This is not meant to imply that an undecidable aspect must arise in all formulations. Rather, the point is: taking *full* account of what the prisoner can deduce can help in understanding the problem (other authors have no doubt recognized this also; it's being emphasized here).

---

[2] The undecidability affects the validity of the prisoner's reasoning but the judge can still make a *choice* for the hanging day.

Thus, there is a view of the hanging problem in which the apparently watertight logic of the prisoner in the statement of the unexpected hanging problem is challenged by an undecidable aspect. It is interesting to observe that the apparently impenetrable logic of the measurement problem as depicted by the sequence of S_i in Section 2 is penetrated by an aspect of uncertainty introduced in the explanation of the measurement problem given in Zurek [1982]. This is based on the interaction of the quantum apparatus with other physical systems (the environment) leading to the avoidance of closed systems such as the S_i in Section 2. Associated with this is not knowing precisely *where*, as opposed to *whether*, the wave function collapse occurs (this is made explicit in Section 2 of Zurek [1981]). In Holtzman [1987], the undecidability of a minimum number of days for satisfaction of the announcement of the unexpected hanging is the crux of the argument. So, in both of these papers, it is uncertainty concerning *problematical transitions* which plays a role in explanation of problems with apparently sound logic but unsatisfying or puzzling consequences. The difficulty associated with each of these transitions is basic to each problem. (Of course, uncertainty is not offered as a general antidote.)

4 CONCLUDING COMMENTS

One may question why there should be parallels between the hanging and measurement problems, especially when one might expect 'weird things' in quantum physics not expected macroscopically. Yet, the fact that there are a number of parallels between the two problems is not surprising considering the nature of the problems. Both problems have strong non-intuitive aspects and seem to be open to all kinds of speculation. Furthermore, the unexpected hanging problem is incompletely specified and is open to various interpretations, formulations, and opportunities for drawing upon various philosophical disciplines—logic, epistemology, etc. (see Margalit and Bar-Hillel [1983], especially the Summary). Thus, there are enough pairs of explanations to obtain parallels. This can make the parallels seem spurious. On the other hand, parallels between such disparate problems may be of particular interest, especially when broader themes surface.

The broader theme that emerged in this note was how the introduction of elements of uncertainty can point the way out of puzzling situations. Some parallel was shown between a proposed resolution to the measurement problem and the undecidability present in a recent view of the unexpected hanging problem. It has been speculated (see Peres and Zurek [1982]) that inability to completely describe the measurement process may be a logical necessity, an inherent undecidability. Making this explicit would, of course, be significant.

ACKNOWLEDGEMENTS

I thank J. G. Loeser and W. H. Zurek for their comments.

JACK M. HOLTZMAN
AT&T Bell Laboratories
Holmdel, New Jersey 07733, U.S.A.

REFERENCES

D'ESPAGNAT, B. [1976]: *Conceptual Foundations of Quantum Mechanics*, 2nd Ed., Reading. Mass.: Benjamin.

HOLTZMAN, J. M. [1987]: 'An Undecidable Aspect of the Unexpected Hanging Problem', *Philosophia*, 17(2), March, 1987, pp. 195–98.

LOESER, J. G. [1984]: 'Three Perspectives on Schrödinger's Cat', *Am. J. Phys.*, 52(12), December, 1984, pp. 1089–93.

MARGALIT, A. and BAR-HILLEL, M. [1983]: 'Expecting the Unexpected', *Philosophia*, 13, October 1983, pp. 263–89.

MELTZER, B. and GOOD, I. J. [1965]: 'Two Forms of the Prediction Pararadox', *British Journal for the Philosophy of Science*, 16, May 1965, pp. 50–1.

PERES, A. and ZUREK. W. H. [1982]: 'Is Quantum Theory Universally Valid?', *Am. J. Phys.*, 50(9), Sept. 1982, pp. 807–10.

PRIMAS, H. [1983]: *Chemistry, Quantum Mechanics, and Reductionism*, 2nd Ed. Berlin: Springer.

ZUREK, W. H. [1981]: 'Pointer basis of quantum apparatus: Into what mixture does the wave function collapse?', *Physical Review D*, 24(6), 15 September 1981, pp. 1516–25.

ZUREK, W. H. [1982]: 'Environment-induced superselection rules', *Physical Review D*, 26(8), October, 1982, pp. 1862–80.

Brit. J. Phil. Sci. 39 (1988), 403–410 *Printed in Great Britain*

REVIEW ARTICLE

Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life

SHAPIN, STEVEN and SCHAFFER, SIMON [1985]: *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life.* Princeton University Press. *xiv*+440 pp. Hardback £43.00. (ISBN 0-691-08393-2)

1 *Introduction*
2 *The Methodology*
3 *The History*
4 *Conclusion*

I INTRODUCTION

Experiment has recently become a central topic of concern in philosophy of science. *Leviathan and the Air-Pump* is a book about experiment. In particular it tells:

> . . . the historical circumstances in which experiment as a systematic means of generating natural knowledge arose, in which experimental practices became institutionalized, and in which experimentally produced matters of fact were made into the foundations of what counted as proper scientific knowledge. (p. 3—All page references will be to *Leviathan and the Air-Pump.*)

The book discusses Robert Boyle's paradigmatic experimental researches using the air-pump. It places them in their historic context and looks at Boyle's experimental/empiricist natural philosophy from the point of view of Thomas Hobbes. Hobbes's *Leviathan* is used as a resource in this history of science. *Leviathan and the Air-Pump* will appeal to two philosophic audiences—historians of philosophy and philosophers of science.

Leviathan and the Air-Pump will appeal to historians of philosophy: it is an account of the confrontation between the natural philosophies of Hobbes and Boyle and of the eventual success of the latter. The account given is a symmetrical account, neither system is prejudged right or wrong but each system is taken on its own philosophic merits. Neither system is completely convincing, nor completely unconvincing. As natural philosophies there seems to be no accounting for the success of one as opposed to the other. The authors, however, take into account the historical circumstances—these natural philosophies are directly related to various social and political views, and these views turn out to have clear bearing on the respective acceptabilities of the two natural philosophies under consideration. In other words, the authors give a social account of the eventual success of one of these natural philosophies.

The book will appeal to philosophers of science in several ways. It will appeal to those with an interest in experiment. It is a study of the institution of the experimental/empirical practice of acquiring knowledge of nature, a practice which is now so widely known and accepted that we seldom think of it as having historic origins. Also the book will appeal methodologically—the method adopted in the study of the historical origins of experimental practice is that of recent sociology of science, it is symmetrical in its discussion of the practice we now routinely accept and those practices which were offered as alternatives. Finally, the book will be of interest to those philosophers of science concerned with the sociology of scientific knowledge, and those concerned with issues in the realism/relativism debate:

> 'Truth', 'adequacy' and 'objectivity' will be dealt with as accomplishments, as historical products, as actors' judgments and categories. They will be topics for our inquiry, not resources unreflectively to be used in that inquiry. (p. 14)

This book, then, though presenting a history of science, reaches tangible conclusions within the domain of philosophy. It establishes a concrete example of the claim that:

> . . . the problem of generating and protecting knowledge is a problem in politics, and, conversely, that the problem of political order always involves solutions to the problem of knowledge. (p. 21)

In this review I first discuss the methodology used by the authors, and then I discuss the historical story they tell.

2 THE METHODOLOGY

The first chapter of *Leviathan and the Air-Pump* explains the method adopted, it explains the way in which the authors will look at this period in history. The method is sociology of scientific knowledge in the Edinburgh tradition. In particular Shapin and Schaffer adopt the symmetry postulate of that tradition. That is, they do not give an account of how Boyle was right and try to explain why Hobbes continued to maintain his wrong position. Rather they maintain that both points of view were credible and viable, and, on this basis, go on to discuss the ways in which each position was developed and defended.

In refusing to take truth and objectivity as given categories to be used unreflectively in discussing this particular episode in the history of science, the authors present the dispute between Hobbes and Boyle more genuinely as a philosophical dispute. By refusing to discuss the dispute in terms of our current judgments about who was right, by refusing to allow those judgments a role in the analysis of the dispute, the dispute is placed in the context of the history of philosophy. The authors adopt a method which attempts to look at their (and our) way of thinking from outside that way of thinking—they attempt to provide a 'stranger's perspective' on our way of thinking. Our routine unreflective judgment is that the experimental/empirical scientific practice advocated by Boyle is the (best) way to generate knowledge. By taking Hobbes's arguments seriously, by accepting his natural philosophy in its own right, it becomes possible to see our routine unreflective judgment from the outside and see the need for its justification. By looking for the justification of this judgment we can see the effort, the labour, that was required to establish Boyle's position.

We are used to thinking of science as an objective study, intellectually independent of wider social or political issues. For us, science is practiced within the scientific community, and the methods used and truths achieved by science are rational and objective practices and products of that community, unaffected (ideally) by extra-scientific social, political or philosophical events. But this way of regarding science, this form of scientific life, is not 'natural' it is not the way things have always been. Instead it is a form of life that had its origins in the 17th Century and had to be worked for, argued for, and defended, because it was not universally accepted as the right way of doing things.

The experimental way of life, the empirical method of science, is now so widely accepted as to be rarely, if ever, subject to any kind of philosophical attack. The experimental form of life is accepted as a matter of routine, it is 'natural'. But this has not always been so—until the 17th Century there was no distinction between 'science' and 'philosophy'. *Leviathan and the Air-Pump* is a study of the origins of this dichotomy we now find so natural, it is a study of the division of natural philosophy. The labour that went into establishing Boyle's experimentalist programme is lost in the gloss of history and the obvious naturalness of that way of doing things is now generally accepted. The authors provide an historical account of how this gloss came about. But a fine-grained historical inspection of Boyle's struggle with Hobbes reveals his labour and the various knowledge-producing tools that Boyle utilized. The method of adopting the 'stranger's perspective' and of looking at Boyle and Hobbes symmetrically enabled Shapin and Schaffer to reveal aspects of this historical epoch that have been hitherto lost through subsequent reconstructions of history.

3 THE HISTORY

Central to the discussion of *Leviathan and the Air-Pump* is Hobbes's *Dialogus Physicus* in which Hobbes responds to Boyle's *New Experiments Physico-Mechanical* of 1660. This dialogue is, in a very real sense, the origin of *Leviathan and the Air-Pump*. It provides the alternative point of view from which Boyle's programme can be inspected—it provides the coherent philosophical alternative to the experimental programme and arguments with which to defend that alternative. It does not accept the form of discourse promoted by Boyle and the experimentalists, but raises questions about the value and justifiability of that form of discourse. It treats the experimentalist programme as another natural philosophy and an inadequate one at that. The *Dialogus Physicus* is a natural philosophical work and meets Boyle's experimental philosophy as another natural philosophy. This, then, is the 'stranger's perspective'. The *Dialogus Physicus* appears in English translation as *A Physical Dialogue of the Nature of the Air: A Conjecture taken up from Experiments recently made in London at Gresham College* in the appendix to the book. For my part, I would suggest that to better appreciate the stranger's perspective adopted in the writing of the book, one should read *A Physical Dialogue* fairly early on, say after having read the first chapter—familiarity with the dialogue makes evident the philosophical nature of the dispute.

Seeing Boyle's programme as one among alternative natural philosophies, we are the better prepared to appreciate the effort that was required to construct and establish his as the correct natural philosophy. One of the major efforts that had to be made was to get people to talk about the right things in the right way—the form of discourse had to be created and defended. One example of the difficulty in establishing the new form of discourse is in Boyle's exclusion of traditional philosophical disputes between vacuists and plenists from the discourse of the experimental natural philosopher. This traditional dispute had proved irresolvable, and Boyle maintained it was futile to pursue it. In its place, Boyle offered a new definition of the vacuum, an operational definition based on the use of the air-pump and his belief that the air-pump removed all (or most) of the air from the receiver. Boyle's definition of a vacuum was a space 'altogether, or almost totally, devoid of air'. In contrast, the traditional definition of a vacuum as a space wherein there is no body at all, allowed the possibility of there being some etherial matter in the space. The nature of this etherial matter and the possibility of its existence afforded a topic of endless discussion and speculation. But this possibility would only be relevant to the experimental programme if it had some experimental consequences. Without such consequences, this possibility became inconsequential to the experimental programme and its discussion was excluded from the experimental discourse. In this way, Boyle attempted to change the discourse, to change the kind of issues being raised. But the difficulty of getting this definition of a vacuum accepted, the *unnaturalness* of this concept, is evident from the fact that Boyle was obliged continually to repeat it in his disputes.

A second major effort that had to be made was in the social organization of the experimental way of life. This is well illustrated by the problems involved in replication and repetition of Boyle's experiments. Within Boyle's programme, the experiments were supposed to exhibit unproblematic self-evident facts that provided the secure foundation of knowledge. But instead of treating the facts as self-evident or self-explanatory, Shapin and Schaffer show how facts depend on social conventions:

> . . . conventions concerning how the knowledge is to be produced, about what may be questioned and what may not, about what is normally expected and what counts as an anomaly, about what is to be regarded as evidence and proof. (p. 225)

The development and establishment of these conventions is shown in a discussion of the reception of a strange phenomena discovered by Christiaan Huygens—the phenomena of anomalous suspension. The general problem of replication is the mutual interdependence of two judgments, whether the apparatus is working and what phenomena it exhibits when it is working. The success of replicating Boyle's experiments depended on the criterion used to determine whether the air-pump was working or not. And the decision as to whether the pump was working or not depended crucially on whether it could replicate Boyle's phenomena. Before an experimenter could test Boyle's results, he had to first accept them as the measure of whether his machine was working well. The only way out of this so called 'experimenters' regress' was through negotiation of the social conventions by which the adequacy of the pumps, and the objectivity of the phenomena produced by them, were judged. The phenomena of anomalous suspension was a phenomena which did not rest comfortably with the pneumatic account of the air-pump, it 'resisted the recognized explanatory competence of the spring of the air' (p. 230). The reception of this new 'fact' is discussed in detail: it affords considerable insight into the negotiations that were involved in this phenomena coming to be accepted as a genuine fact. The authors conclude that:

> The establishment of a set of accepted matters of fact about pneumatics required the establishment and definition of a community of experimenters who worked with shared social conventions: that is to say, the effective solution to the problem of knowledge was predicated upon a solution to the problem of social order. (p. 282)

These attempts to reform the discourse of natural philosophy and to organize the social basis for the experimental way of life were, and were seen to be, integrally part of larger issues. Hobbes saw talk of the vacuum as a danger to society. He argues that talk of a vacuum will give rise to dissention within the State and eventually to civil war. In *Leviathan* Hobbes attacked the vacuum on political grounds.

> The vacuum Hobbes attacked was not merely absurd and wrong, as it was in his physical texts; it was *dangerous*. Speech of a vacuum was associated with cultural resources that had been illegitimately used to subvert proper authority in the state. (p. 91)
>
> In Hobbes's view the elimination of vacuum was a contribution to the avoidance of civil war. The dualist ontology deployed by priests spoke of existents which were not matter: this made men 'see double' and resulted in the fragmentation of authority which led inexorably to chaos and civil war. (p. 18)

Boyle countered this argument with the claim that Hobbes's system subverted the Christian religion. Hobbes believes in a material, or corporeal, God. He also believed in a world completely full of matter, a plenum. Boyle argued that it is difficult to conceive how such a corporeal God can have access to a world already full of matter. There must be some vacuum in the world for God to gain access to it. Thus Boyle suggests that to deny there is any vacuum in the world amounts to an assertion of atheism:

> Boyle invited natural philosophers to reject Hobbes's physics because it conduced to irreligion, and he suggested that his adversary's civic philosophy and theology could be invalidated if it were shown that his physics was unsound. Thus Boyle's defence manifested the typical seventeenth-century network of calculation whereby criteria used to evaluate good religion were built into the evaluation of good natural philosophy and *vice versa*. . . . The same conceptual resources that, according to Boyle, threatened a proper idea of the Deity and his attributes simultaneously threatened the perceived integrity of the machine whose workings were the key to proper natural knowledge. Boyle's ideological assault on Hobbes was, therefore, an integral part of his defence of experiment and of the engine that was its powerful and emblematic device. (p. 207)

4 CONCLUSION

In *Leviathan and the Air-Pump* few, if any, general, testable hypotheses are being made about the relations between social history and the success or failure of substantive philosophical or scientific claims. This book is a particular study of a particular epoch in the history of science. It documents the development and establishment of the experimental way of life and the scientific way of generating knowledge as we now know them. It shows that development and establishment to have been a result of contingent aspects of a particular historical epoch. Neither does the book make any speculative projections about how knowledge would have been if things had been different—the only remark the authors make is that:

> Given other circumstances bearing upon that philosophical community,

> Hobbes's views might well have found a different reception. They were not widely credited or believed—but they were *believable*; they were not counted to be correct—but there was nothing inherent in them that prevented a different evalution. (p. 13)

However, it is not generally thought that history is readily regimented by general, testable hypotheses that can be used to support counter-factual speculations. Criticism of social history of science on the grounds that it affords no such general claims is mistaken criticism.

But the authors do draw historical conclusions. They do not claim to answer the question 'Why did Boyle win?' because they feel that 'an unbroken continuum between Boyle's interventions and twentieth-century science is highly unlikely' (p. 341). But they do claim to have accounted for the origins and social organization of the experimental/empirical basis of scientific knowledge:

> In this book we have examined the origins of a relationship between our knowledge and our polity that has, in its fundamentals, lasted for three centuries. (p. 343)

The decision to adopt Boyle's programme over that of Hobbes was not a decision that could be made rationally—the programmes differed on what they considered good reasons for accepting something as knowledge. Both Boyle and Hobbes offered their programme, their natural philosophy, as a part of a more general solution to the political problems of Restoration England. But in this there was no further hope of resolution—their political programmes were as diverse as their philosophical methodologies and their natural philosophies. In a sense then the two programmes are incommensurable—they do not have a common ground on which to settle disputes.

Shapin and Schaffer account for the success of Boyle's natural philosophy on the basis of political considerations:

> We . . . consider the issues that bore on the way Hobbes's and Boyle's schemes were assessed in the 1660s. This demands an outline of the political and ecclesiastical context of the Restoration. . . . These political considerations were constituents of the evaluation of rival natural philosophical programmes. (p. 283)

And they do this accounting on the basis of social history:

> . . . the contest among alternative forms of life and their characteristic forms of intellectual product depends upon the political success of the various candidates in insinuating themselves into the activities of other institutions and other interest groups. He who has the most, and the most powerful, allies wins.
>
> We have sought to establish that what the Restoration polity and experimental science had in common was a form of life. (p. 342)

What the authors establish is a specific claim about the history of science, the development and institution of the experimental way of life. Sociology of scientific knowledge takes it as a methodological principle that social accounts of the acceptance of natural philosophies or scientific theories can always be given. The accomplishment of Shapin and Schaffer is to give such an account of a way of life and of knowledge generation which has remained virtually uncontested for three centuries.

RICHARD C. JENNINGS
University of Cambridge

Brit. J. Phil. Sci. 39 (1988), 411–414 *Printed in Great Britain*

REVIEW ARTICLE

How Experiments End: Galison, Peter [1987]

ALLAN FRANKLIN. University of Chicago Press. *xii* + 330 pp. $39.95 (cloth), $14.95 (paper)

One of the most interesting and important trends in the history and philosophy of science has been the recent work on experiment. Most philosophy of science, and sometimes even history of science, either neglects experiments—how they are done and what role they play—or treats their results as unproblematical. Peter Galison's *How Experiments End* is a major contribution to the growing body of work that is correcting that view. Galison provides excellent histories of three experimental episodes: the measurement of the gyromagnetic ratio of the electron, the discovery of the mu meson, or muon, and the discovery of weak neutral currents. These studies of actual experiments will provide valuable material for both philosophers and historians of science and Galison's own thoughts on the nature of experiment are extremely important.

One of these is that, contrary to the prevalent view of theory dominated science, experimental practice and instruments often persist across major changes in theory, and thus provide continuity across these conceptual changes. Thus, the experiments on the gyromagnetic ratio of the electron spanned classical electromagnetism, Bohr's old quantum theory, and the new quantum mechanics of Heisenberg and Schrodinger. As Hacking pointed out, and Galison reinforces, experiment often has a life of its own. This is not to say that Galison neglects theory, but rather that he offers new insights into the way in which experiment and theory interact. This view of experimental continuity is tied to Galison's concept of experimental traditions, in which scientists develop skill in using certain types of instruments and apparatus and will therefore regard particular kinds of evidence as most convincing. In particle physics, Galison discusses the traditions of visual detectors such as the cloud chamber and the bubble chamber in contrast to the electronic tradition of Geiger counters, scintillation counters, and spark chambers. Scientists within the visual tradition tend to prefer 'golden events' that clearly demonstrate the existence of a phenomenon, such as Anderson's photograph of the positron. In the electronic tradition, statistics tend to be more important than single events. In the weak neutral current episode a large number of events were needed so that the real effect could be separated from neutron background.

Galison also emphasizes that the decision to end an experiment, to accept a result as valid, is very complex. In large groups, certain pieces of evidence will be more convincing for some group members than for others. In the Gargamelle experiment on weak neutral currents, several group members, particularly Perkins, found the single photograph of neutrino-electron scattering particularly important. For others, the difference in the spatial distribution between the observed neutral current candidates and the neutron background was decisive. In contemporary high energy physics, the extremely large experimental groups may contain members from both the visual and electronic traditions. It may already be the case that some of the most interesting discussions and arguments concerning the validity of experimental results occur within these large experimental groups and are not revealed in the published work. One of the interesting features of this book is that it makes such arguments available.

The theoretical presuppositions of the experimenters may also enter into the decision to end an experiment. Einstein and de Haas ended their search for systematic errors when their result for the gyromagnetic ratio, $g=1$, agreed with their belief that magnetism was due to orbitting electrons. As Galison points out, the search for systematic effects is both subtle and difficult. If one looks at the history of measurements of η_{+-}, the CP violating parameter in K° decay, one finds that the world average of measurements before and after 1973 differ by eight standard deviations, an extremely unlikely result if the two sets of measurements are both valid measurements of the same quantity. It is clear that systematic effects have been overlooked by some of the experimenters. A reasonable point at which to stop looking for such systematic effects is when the result agrees with previous measurements or when it agrees with existing theory. Theory may also influence what is considered to be a real effect, demanding explanation, and what is considered background. Galison shows that part of the discovery of the muon involved the calculation by Oppenheimer and Carlson, which showed that showers were to be expected when electrons passed through matter, and that what needed explanation was the behavior of penetrating particles, later found to be the muon. Theory may also give both the size of the effect expected as well as the size of the expected backgrounds. This will indicate whether or not an experiment is feasible. Galison does not say that theory determines the results of an experiment, but he does argue, very persuasively, that both the theoretical and instrumental commitments of the experimenters help to determine the alternatives considered.

In this discussion, Galison also emphasizes that the elimination of background that might simulate or influence a result is not a peripheral activity, but one that is central to the experimental enterprise. In the Gargamelle experiment, events that apparently demonstrated the existence of neutral currents could also be produced by neutron background. A considerable part of demonstrating the existence of neutral currents consisted of showing that the observed events could not be due to such neutron background. In E1A, the other neutral current experiment, the group believed, for a time, that they had not observed the currents. This was because hadrons were punching through the muon shield, simulating muons, and eliminating valid neutral current events. It was only after a complex Monte Carlo calculation was performed, and checked, that this effect was eliminated. Because the group had found evidence earlier for neutral currents, this led to some wags to remark that 'alternating' weak neutral currents had been found.

Galison emphsasizes that it is not whether or not theory enters into the decision to end an experiment, but where and how it enters. He distinguishes between long, medium, and short term theoretical commitments. Thus, in the neutral current experiments the long term commitment was the unification of weak and electromagnetic forces, the medium term was gauge theories, and the short term was the particular Weinberg-Salam theory of electroweak forces. There are similar experimental, or instrumental, commitments. Galison gives instrumental type, specific device, and a particular experimental run as the long, medium, and short term experimental commitments, respectively. These, too, will influence the decision to accept a result and end an experiment.

Recently, there has been a continuing argument between those who regard science as a reasonable, dare one say rational, activity and those who view it as merely a social construction by scientists—between those who consider evidence important and justified and those who argue that it is just socially accepted practice, with no deeper justification. For example, social constructivists argue that the dominance of theory will always guarantee that experiment will fit theory, and thus give rise to global incommensurability between theories. Galison argues that there are three difficulties with this approach. The first is that although experiment may not logically compel a particular theoretical conclusion, it may provide very persuasive, if not convincing, evidence. The second point is that the social constructivists exaggerate the flexibility of theory. Although it may be possible in principle to adjust a theory to fit any particular piece of evidence, in practice the mathematical and physical constraints on a theory are not easily dismissed. What is logically possible may not be physically plausible or interesting. A third point is that there are constraints on experimentalists' conclusions imposed by the skills and techniques of their work. One might also add that there may be very good arguments to support the validity of a particular experimental result. This position fits in with Galison's view that experimental techniques, instruments, and results persist through major theoretical change. Galison recognizes, and documents in detail, that the decision to end an experiment is a social process, but he argues persuasively that it is based on evidence that 'will hold up in court'. In contrast to Pickering, who, in his study of the weak neutral current experiments, argued that the discovery was only a change in the interpretive practice of physicists, with the implication that the conclusion will change if the practice changes, Galison shows how complex, difficult, and ultimately reasonable it was to conclude that the currents existed. He cites David Cline, one of the E1A expermenters, who wrote in an internal memorandum, "At present I don't see how to make these effects go away." Galison notes that although theory can tell you what to look for, and what experimental cuts, selection criteria, to make, it cannot guarantee that the events will be seen. The world does have an effect on what is observed.

There is one minor quibble I have with Galison on this point. He correctly emphasizes that it is the judgement of the scientific community that is really at stake, and that the evidence that is used to persuade those outside the experimental group may differ from that which persuades a group member. He has concentrated on arguments internal to the group and has, I believe, slighted the published arguments. I wish these public arguments had been given a little more emphasis.

In this review I have concentrated on some of Galison's points that will, and should, be of real interest to philosphers of science. His previously published studies of these episodes, which are included in this book in modified form, have already given us much of value, but in placing them in a deeper and more general context he has enhanced their value. I have not done justice to the wealth of historical detail presented or to the excellent writing. Galison understands the physics and tells his stories in an exciting way. Even though I had read the papers, I found myself eager to find out what happened next, a trait I usually find more often in fiction than in the history of science.

If there is one criticism I have of this excellent book, it is that Galison tends to slight the more traditional philosophical questions concerning experiment, evidence, and theory. Perhaps this is understandable in a book that gives us a new and valuable way of looking at these issues, but I would have liked to see a little more epistemology, for example. Galison argues that in these episodes the experiments were progressively more direct and that the results achieved increasing stability. Directness is achieved by eliminating background and stability by varying experimental conditions and changing methods of analysis. I agree with Galison that these are good arguments, and that they are part of what we intuitively recognize as good science, but why they are, and should be, good reasons to believe a result is not dealt with extensively. This is, however, a minor criticism. Galison has given both philosophers and historians much to think about. I strongly urge you to read this book.

Brit. J. Phil. Sci. 39 (1988), 415–419 Printed in Great Britain

REVIEW

EVANDRO AGAZZI (Ed) [1986]: *Epistemologia: La filosofia della scienza in Italia nel '900*

Franco Angeli s.r.l., Milano, pp. 1–512
Price: L.40'000.-, Iva incl.

ANDRÉ MERCIER
University of Berne (Switzerland)

Eighteen authors have contributed to this huge review of the Philosophy of science in Italy during the 20th century. Their contributions cover the origins of modern philosophy of science at the turn of the 19th to the 20th century, the philosophy of science during the first half of the 20th c. (until 2nd World war), the 'new philosophy of science' that arose after the war, and contemporary trends in the field. All this is framed between a critical Introduction by AGAZZI himself and an Appendix containing the list of reviews and main publications, as well as names of Journals as produced in Italy, analytically and critically presented.

Subject-matters as positivism and pragmatism, foundation of mathematics, intuitionism, logistic and artificial languages, idealism, the interpretation of relativity theory, quantum physics, the particular period of transition and renewal lasting from the First till the Second world war, neo-positivism, Marxist tendencies, history versus philosophy of science, the special philosophies of mathematics and of physics, of cybernetics and systems, of social science, objectualistic tendencies, and Popperianism in recent years, are dealt with. The many '-isms' included in that list are typical of what has been done and how it has been done in all these hundred years in Italy.

Indeed, by looking merely at the title of the book, a naive reader might expect to find a series of reports on the special contributions made by Italian savants and thinkers to the philosophy of science, summarizing their works and listing their main results and advances. If this were the sole expectation of the reader, he would be disappointed. For, very little of that is explicitly present in the book.

Surely, to begin with, representatives of the older generation active in the field of philosophy of science like the author of the present review have of course in the late twenties and the thirties of this century heard of names like those of PEANO, ENRIQUES, FANTAPPIÉ and others in reference with what was then described as the Crisis of foundation of mathematics, and they had eventually read and studied their papers and books. Therefore they might think that e.g. the main results attained by such authors would be described and judged according to their importance in the light of the present situation fifty years or more later. Or one might wish to know what are the specific contributions of more recent authors to the progress of the philosophy of the various specific sciences ... However, little of that constitutes the contents proper of this big book. Rather, many of the contributions are papers on what might be called partly the history, partly the critique of the polemics between schools of thought with special reference to the philosophy of science. Still, for all that, some contributions are more precisely conceived as a clear presentation of what scholars have contributed to the clarification of problems of the philosophy of the various scientific fields.

So the question arises, why the book under review more or less inevitably presents itself as it does. The reason seems to be found in the very nature of traditional philosophizing in Italy still very much practised today: Philosophers in that country do for a considerable part live on, and love, philosophical controversy. This may be both their strength and their weakness. Weakness, evidently, because if one persists in quarrelling with others about which is the right stand-point, he will easily miss the train and remain rather unproductive within the actual disentanglement of specific problematics at hand. Strength, because it makes the scholars extremely sharp in the art of rhetorics and argumentation.

This may explain why the development of philosophy in Italy, even in a field as precisely delimited as the philosophy of science, has consisted in a long and weary controversy between the -isms, much more so than it has been the case, especially before the Second world war, in the rest of the Continent. Of course, such controversies did also take place, e.g. in France, Holland and elsewhere between extreme tendencies like neo-positivism or intuitionism, but there, the situation was less polemic and disputatious, for the scholars would prefer to write matter-of-fact papers.

It may also be due to two circumstances. One is that, during the 18th and 19th centuries, Italian philosophy, especially as it was concerned with the advance of the sciences, had declined. Hence, on the turn between the 19th and the 20th century, when not only professional philosophers but also and advantageously scholars in the specific fields—especially mathematicians and physicists—began to write on the philosophy of science (mainly in France and Germany), this very kind of philosophy developed on a scene where standpoints soon began to oppose one another. And since controversy seems to be a favourite occupation among Italians, this particular kind of philosophizing developed. It was nevertheless luck that mathematicians of rank were among the protagonists (PEANO and others have already been mentioned and a number of others could be named. The only name still more often quoted in the book, yet for a more recent period, is that of LUDOVICO GEYMONIAT, also very much interested in mathamtics, a 'neo-illuministic libertarian' as he is called at one place in the book).

Another circumstance may reside in the fact that philosophy, generally speaking, was, until rather late in the 20th century, very much dominated in Italy by two personalities of high status: CROCE and GENTILE, who were outstanding protagonists of an idealism which not simply combatted scientism and positivism, but claimed to explain science as a reductive and pragmatic type of cognitive activity. For CROCE, science was outside philosophy. Backed by historicism, idealism in Italy has been qualitatively and quantitatively very much represented until our day, but it has also gone through an evolution as regards the consideration of science, since GUZZO for instance insisted upon the humanistic value of the sciences, or SPIRITO attributed to them an active rôle in the regeneration of philosophy.

After the last war, everything has changed in the whole world. Also in Italy, one had to adapt philosophic reflection to a totally new cultural, political and epistemological situation. The strongest impulse came from L. GEYMONIAT, however, as late as the sixties. Nevertheless, in the thirties and forties already, a scientific phenomenon had started in Italy after many decades of relative unproductivity and stagnation: a school of physics emerged of great quality and intense activity, mostly concentrated on theoretical quantum physics and mainly situated at Rome, yet also with ramifications elsewhere and in the other fields of that science (relativity theory, astrophysics and the like). The central figure there has been ENRICO FERMI. Even though he cannot be said to have been a 'philosopher of physics' in the proper sense himself, this emergence of physics which he symbolized has lately been accompanied by the appearance of a number of physicists (also astronomers etc.) interested in, and working specifically on problems of the philosophy of physics, even more so than it is the case in other countries of comparable size. This has some analogy with what had happened in the early years of the century with reference to mathematics. For, if today there is, so to speak, no problem (or 'crisis') of the foundation of mathematics left, there is—since the beginning of the BOHR-EINSTEIN controversy—still with us, a problem (or 'crisis') of the foundation of (theoretical) physics. In contradistinction with older days, however, Italian scholars today refrain in that respect from mere controversies and are mainly concerned with matter-of-fact questions.

Another field, whose importance has been recognized and has increased among Italian scholars recently, is the philosophy of social science and of the humanities as sciences. There of course, one may expect adepts of Marxism and the like (GRAMSCI and others) to be particularly active. Where this is the case, however, they are of the traditional disputatious kind rather than true to the real authentic problematics.

The relationship between the so-called empirical sciences and metaphysics is being approached nowadays in a way rather free from prejudices, as is known to exist, e.g. within analytical and similar schools abroad. This problematic may become more and more important in the coming decades. Here, AGAZZI himself is contributing important material, especially because he is trained not only in traditional philosophy, but also in formal logic and the history of the (exact) sciences. His impact on the consideration of science *versus* ethics is also of considerable bearing for the future.

It appears preferable to refrain from quoting a lot of names, for there are so many cited in the book, that a choice might sound as uttering a kind of judgement of value and this is not possible to do among living philosophers. If at the dawn of the century the interest was mostly in the foundation of mathematics, it also had views on Darwinism and, very specifically yet oddly as it appears today, on the possibility and the expectation of successfully constructing artificial languages (there, the friendship between PEANO and the French mathematician COUTURAT degenerated into an estrangement; in our days, nobody thinks anymore of the applicability of an artificial world language, yet there was a time when this was considered a main possibility for the achievement of universal understanding and peace!).

The book does also give some information on the work done in the field of history of science, especially the relationship between the history and the philosophy of science.

The abundance of papers and books published in Italy on the various topics of the philosophy of science is amazing. Apart from translations of standard foreign works, many more have been translated from French, English, German etc. Of course, Italian scholars are ordinarily very good at French as well as at English or German and would not really need these translations, but the fact remains and is a proof of the lively interest taken in foreign activity in the field. They, themselves, write of course mostly in their own language, but quite a few do also publish—mainly articles—in foreign Reviews and with foreign Publishing companies. Innumerable publishing companies are spread throughout Italy; many have epistemological titles on their lists (we quote here only the most renowned ones: Armando in Rome, Feltrinelli and Angeli in Milan, Il Mulino in Bologna, Cedam in Padua, Bombiani, Boringhieri, Mondadori, Einaudi, Zanichelli and others). The number of Journals where papers on the philosophy of science are published is amazing; from the celebrated *Rendiconti di Palermo* (and the other Academies of course), the more famous journals *Scientia* and *Logos*, till the most recent *Epistemologia*, we would count at least a dozen journals. Some more or less extensive and/or critical reports have been published—in Italy or abroad—on developments of the philosophy of science in Italy, especially by FILIASI-CARCANO, SELVAGGI, SOMENZI, AGAZZI, PERA . . .

The book under review is rather difficult, at places tedious, to read. Several contributions are written in a cumbrous Italian. The easier ones are mostly written by scientists. The old French man of letters BOILEAU said that *ce qui se conçoit bien s'énonce clairement.* Yet be that as it may, both EDMUND BURKE and ALBERT SCHWEITZER have drawn the attention of philosophers upon the fact that the clearer a text, the poorer it is in contents!

The London School of Economics and Political Science

LAKATOS AWARD IN PHILOSOPHY OF SCIENCE

This Award is for an outstanding contribution to the philosophy of science in the form of a book published in English. The Award is endowed by the Latsis Foundation in memory of Imre Lakatos and is administered on behalf of the London School of Economics by a committee consisting of the Director of the School or his deputy as chairman, and Professors Hans Albert, Adolf Grunbaum, Alan Musgrave and John Watkins. The Committee makes the Award on the advice of an independent and anonymous panel of selectors. The value of the Award is £10,000. In any given year the Award may be shared or no Award made.

To receive an Award a successful candidate must visit the School and deliver a public lecture.

Candidates must be nominated by at least three people of recognised professional standing. Nominators should give their grounds for the nomination and indicate the candidate's age, since a preference may be given to a younger scholar. Three copies of the book should, if possible, be sent to the address below. All communications should be marked 'Lakatos Award' and addressed to:

The Secretary,
The London School of Economics,
Houghton Street,
London WC2A 2AE,
England.

The 1989 Award will be for a book published during the last six years (bearing an imprint from 1983 to 1988 inclusive). The closing date for nominations is 15 April 1989.

Teaching
Philosophy

Contents	March, 1988 Vol. 11/1
Thomas Carson	*"Who Are We to Judge"*
Jeffrey Gold	*Bringing Students Out of the Cave*
Thomas Wartenberg	*Teaching Women Philosophy*
Thomas Mappes	*What is Personal Ethics, and Should We be Teaching More of It?*
James Garson	*Heuristics for Proof Finding in Formal Logic*
Kevin O'Neill	*Undergraduate Apprentices in Introductory Philosophy*
Phillip Cummins	*Sophistical Sam's Sad Condition, A Fable*
Reviews	***New Publications***

1988 ***Subscriptions*** *(4 issues): $18.50 individuals; $42 others. Add $4 for postage outside the USA. Send orders to the Philosophy Documentation Center, Bowling Green State University, Bowling Green, OH 43403-0189.*

Edited by Arnold Wilson, University of Cincinnati, Cincinnati, OH 45221-0206.

REVUE INTERNATIONALE DE PHILOSOPHIE

Editor: Michel MEYER
143, av. A. Buyl, 1050 Brussels, Belgium

Each number is devoted to a particular movement, a particular philosopher, or a particular problem.

We publish 3 issues annually. Articles are written in English, French, German, or Italian.

Coming numbers

- Locke
- Kant and Practical Reason
- Artificial Intelligence
- Heidegger
- Wittgenstein

Our last issue was devoted to *Schopenhauer*

(articles by D. W. Hamlyn, D. E. Cartwright, W. Schirmacher and K. Boullart)

TOWARD A NEW PHILOSOPHY OF BIOLOGY

Observations of an Evolutionist

Ernst Mayr

"[A] book to be developed, to be argued with, a book whose margins should be filled with excited scribblings. Philosophy of biology does not stop here, but as Mayr's title indicates, this is a place to start."

—Philip Kitcher, *Nature*

Belknap £27.95 cloth/
ISBN: 0 674 89665 3

Harvard University Press
126 Buckingham Palace Road, London SW1W 9SD

Brit. J. Phil. Sci. 39 (1988), 421–439 *Printed in Great Britain*

*Hattiangadi's Theory of Scientific Problems and the Structure of Standard Epistemologies**

MARCO GIUNTI

---

1 *Introduction*
2 *Hattiangadi's Historical Theory of Problems (an improved version)*
3 *Critical Comments*
    3.1 *Criticism of Hattiangadi's Historical Theory of Problems*
    3.2 *Criticism of Hattiangadi's Logical Theory of Problems*
    3.3 *The Relation Between the Logical and the Historical Theory*
    3.4 *Criticism of Hattiangadi's Methodology*
4 *Concluding Remarks: Hattiangadi's Theory and the Structure of Standard Epistemologies*

---

1 INTRODUCTION

Hattiangadi's two papers on the structure of problems (see [1978] and [1979]), together with Laudan's book *Progress and its Problems*,[1] represent the most complete attempt thus far to ground a whole epistemology on the basic notion of scientific problem.

The most interesting feature of Hattiangadi's system is that both the evaluation of scientific theories and the rules for their acceptance or rejection require the consideration of the historical structure of scientific problems. Hattiangadi has also proposed a logical analysis which is intended to provide the 'unit structure' of all intellectual problems. According to this analysis, all problems arise from logical inconsistencies, while scientific (or deep) problems also have an additional historical structure.

Although Hattiangadi's approach is novel on many significant points, it surprisingly turns out to be an instance of standard epistemology. By this term, I mean any view of science which primarily focuses on the process of theory-choice, and which tries to elaborate a set of evaluation-concepts and of rules for the acceptance or rejection of hypotheses. Standard (or traditional) epistemologies accept the basic postulate:

---

* I wish to thank Prof. Nozetta Koertge for her comments, suggestions, and criticisms.

[1] *Progress and its Problems* (Laudan [1977]) had been published before Hattiangadi's papers, but, as it results from Hattiangadi's autobiographical remarks, his theory was developed independently (see Hattiangadi [1978], pp. 346–7, and p. 355: 'I had for many years . . . their history').

(*TC*) (*under ideal conditions*) *scientists choose theories by following methodological rules, which are based on evaluation-concepts.*

This is a very usual claim, so entrenched in our epistemological beliefs, that it is often taken as the starting point for understanding the whole dynamics of science. According to this view, the problem of any epistemology consists in establishing what the evaluation-concepts, and the related rules for theory-choice are. An epistemology of this kind is not necessarily normative, since the rules governing the process of theory-choice might be taken as not rationally justifiable. It is my contention that it is reasonable to challenge (*TC*). I will argue that (*TC*) is not the only central fact of science, and that, as a consequence, the basic problem of epistemology should be revised.

The second section gives a brief exposition of the motivations and content of Hattiangadi's theory of scientific problems. In the third section, five main criticisms will be presented:

1. the historical theory of problems turns out to be conceptually weak, because some of its main assumptions have an arbitrary character;
2. the view that all problems are logical inconsistencies is not tenable;
3. contrary to Hattiangadi's claims, his analysis of the logical structure of problems does not explain how a theory, T_1, can have some problems which are solved by another theory, T_2, which, in turn, has some other different problems;
4. Hattiangadi's theory of the historical structure of problems cannot be regarded as an extension of his analysis of the logical structure. The two theories are independent, and possibly incompatible;
5. Hattiangadi's new methodology, in both its normative and descriptive aspects, leads to pradoxical results.

These criticisms show that Hattiangadi's detailed proposal for a problem-based approach to epistemology is not successful. This conclusion, however, leaves open the following question: if we are interested in constructing an adequate theory of the development of science, is it reasonable to assume that the concept of scientific problem should be one of its central notions? The answer to this question will be positive, and in Section 4 it will be argued that a problem-based view of science naturally leads to a new epistemological approach, more adequate and comprehensive than the standard one.

2 HATTIANGADI'S HISTORICAL THEORY OF PROBLEMS (AN IMPROVED VERSION)

Hattiangadi presents his theory of problems as a response to two difficulties of Popper's epistemology, which Hattiangadi interprets as a Socratic view of knowledge. According to this interpretation, science does not (and cannot) yield knowledge in the traditional sense (episteme). However, scientists continue to seek knowledge. It is then very natural to ask:

(a) why should we continue to seek knowledge, if we cannot reach it?
(b) If we never have any knowledge, how it is possible that, sometimes, we recognize a certain hypothesis as what we were looking for?

Hattiangadi claims that we can easily solve these difficulties, if we accept a problem-based view of science. The answer to the first question is that we seek knowledge '*because we must, in order to solve problems which arise in our beliefs*' (Hattiangadi [1978], p. 351). The reason why we *must* solve problems is not theoretical, but, ultimately, practical. In fact Hattiangadi maintains that all problems are logical inconsistencies, and, since from a contradiction anything follows, a set of inconsistent beliefs cannot have any practical utility. This [illegible] to question (a), besides being very persuasive and clear, has the [illegible]tage of not using Popper's concept of verisimilitude, which, in turn, is [illegible]ted by at at least two other difficulties (see Hattiangadi [1978], pp. 346–7, [illegible] [1979], pp. 69–70: 'Popper's answer . . . intellectual activity?').

The answer to question (b) is that problems pose some desiderata for a good solution, so that we can recognize whether a certain hypothesis meets these requirements.

According to Hattiangadi, we must distinguish between two main categories of problems. All intellectual problems share the same logical structure, that is, they are contradictions. However, some problems have a further characteristic, *depth.* Scientific problems fall under this category. Their depth is due to the fact that they arise in theories which are embedded in intellectual traditions. Consequently, to understand scientific problems, we have to consider not only their logical structure, but also their historical structure. I am going to present below Hattiangadi's theory of the historical structure of problems.

Competing *intellectual traditions* are lineages of theories and problems with a common origin. All the competing traditions which stem from the same origin form an *intellectual debate.*[2]

The origin of an intellectual debate is constituted by a *problem* which arises for a *theory.* This problem can be *solved* in different ways, by different theories. Each of these different solutions may be the starting point of a distinct intellectual tradition.

Let us consider Figure 1. T_0 is the 'founding theory', and P_0 is the original problem of the debate. This intellectual debate has two distinct traditions, whose origin is, respectively, T_1 and T_2. Any T_i gives a different solution to P_0.

---

[2] Hattiangadi uses 'intellectual tradition' and 'intellectual debate', more or less, as synonyms, and they correspond to what I call here 'intellectual debates'. My 'traditions' are called, by Hattiangadi, 'subtraditions', and, sometimes, 'main lines of a tradition', or 'lines'. The problem with Hattiangadi's terminology is that he does not draw a clear distinction between subtraditions (or main lines) and lines. This distinction becomes crucial when he introduces the concept of dominance of a main line upon another main line.

Origin of an intellectual debate

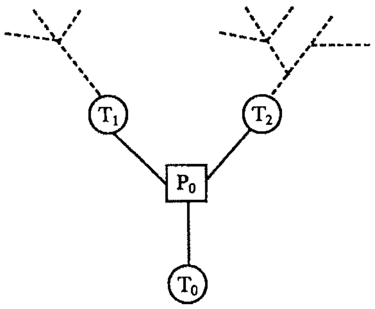


FIGURE 1

In turn any T_i will typically give rise to other problems, some of which solved by other theories, and so on. The debate in Figure 1, at a later might look as in Figure 2.

The intellectual debate in fig. 1 at a later stage

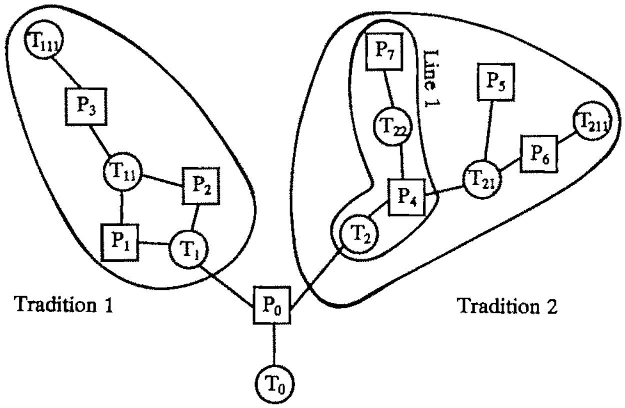
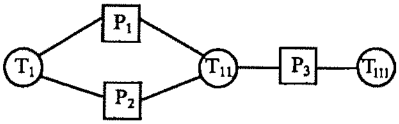


FIGURE 2

There are two remediable, but intriguing, *formal* problems with Hattiangadi's tree-like representations of intellectual debates. One of them concerns the concept of *line of thought* in a tradition of an intellectual debate. The other regards the concept of *discriminatory problem*. These two notions are very important for Hattiangadi's epistemology, since they will be employed to define a set of evaluation-concepts, which allow the appraisal of traditions. These concepts, in turn, are the basis of Hattiangadi's methodological rules for the acceptance or rejection of theories. For this reason, I am going to discuss in some detail Hattiangadi's basic notions of line and discriminatory problem.

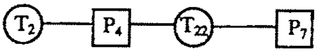
Tradition 1 in Figure 2 has one *line of thought*, the line

[0]



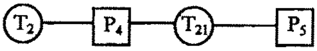
How many lines has tradition 2? One line certainly is

[1]



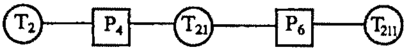
The problem is with the other branch of the tradition. Should we count this as a unique line or as two different lines? Hattiangadi is not clear on this point. The simplest answer would consist in taking

[2]



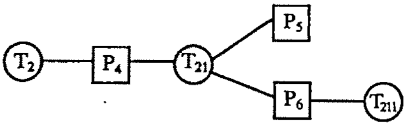
and

[3]



as two different lines. If we make this choice, we have that [2] has one open problem (P_5), while [3] has no open problem. Consequently, tradition 2 is, on the whole, as successful as tradition 1, which also has one line with no open problem. But, according to Hattiangadi, in a case like this we should conclude that tradition 1 *dominates* tradition 2 (see Hattiangadi [1979], p. 61, schema 7). On the other hand, a tradition dominates the other if, and only if, it has at least one line with no open problem and all the lines of the other tradition have at least some open problems. Consequently, tradition 1 does not dominate tradition 2. We have a contradiction, and we thus must conclude that [2] and [3] are not different lines. We then have one line:

[4]

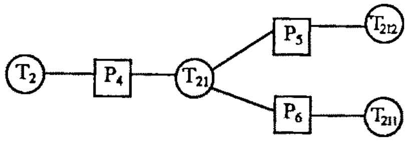


line 2 tradition 2 in fig. 2

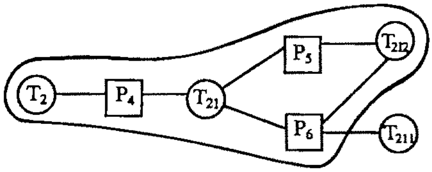
How many *open* problems does this line have? The most natural answer is one. Nevertheless, according to Hattiangadi's theory, they should be two, P_5 and P_6, even though P_6 is *solved* by T_{211}. The basic idea is that P_6 is still open because it is not solved by a theory which also solves P_5. Hattiangadi does not explicitly draw this distinction between solved, but still open, problems, and definitely solved, or *closed*, problems in a line. This distinction is, however, implicit in his claim that 'a subsequent problem includes within its structure all the earlier problems on that line' (Hattiangadi [1979], p. 55).[3]

What would it happen if line [4] developed as, respectively, [5] or [6]?

[5]



[6]



line 2 of tradition 2 in fig. 2 at a later stage

The whole [5] would still count as a unique line, with two open problems, P_5 and P_6.

In [6], T_{212} has solved both problems of T_{21}, P_5 and P_6. As a consequence, T_{211} has been *expelled* from the line, which now has no open problem.[4]

Let us now define: a line is *closed* iff it has no open problem; a line is *open* iff it has at least one open problem. We can finally define the concept of *dominance* of a tradition over another tradition. Tradition TR_1 dominates tradition TR_2 (in debate D) iff at least one line in TR_1 is closed, and no line in TR_2 is closed (see Hattiangadi [1979], p. 61, schema 7).

---

[3] This property of lines in a tradition might be called the *conservation-property*. No previous problem is ever lost along a line. If some problem which arises in theory T_i of a line is not solved together with the other problems which arise in T_i, *all* the problems of T_i remain open, until a subsequent theory, T_j, closes the line. When T_j closes the line (as in [6]) all problems of T_i become solved problems of T_j. Note that the closing of the line might not happen immediately, as in [6]. It might occur much later in the development of the line. The effect would however be the same.

There is another important property of lines. This can be called the *non-recurrence property* (see Hattiangadi [1979], p. 54: 'But it is . . . that line'). In one line, no solved problem reappears in a successive theory. There is a minor difficulty with this property. Schema 5 in Hattiangadi [1979], p. 57, has one line which does not satisfy the non-recurrence property. Specifically, problem B occurs twice in the same line. I take this to be just a slip of the pen.

[4] I must here point out that the concept of *expulsion*, and the elaboration of the notion of line based on the distinction between *solved* and *closed* problems, are not Hattiangadi's. His theory simply lacks a clear notion of line, which, however, is crucial for introducing the concept of dominance, and standing, of a tradition with respect to another tradition. I maintain that the interpretation I give here grasps Hattiangadi's intuitions. However, the main critical points I will make later are independent from this interpretation. This interpretation has the only function to make sure that Hattiangadi's theory *can* be elaborated so that it is, at least, formally adequate.

The relative *standing* of two traditions is the nearness of one to being dominated by, or dominating, the other one (see Hattiagadi [1979], p. 61).[5]

A tradition *TR* (in debate *D*) is *viable* iff no other tradition in *D* dominates *TR*.

The *value* of a solution of a problem, in debate *D*, is given by the difference that it makes to the standing of the different traditions in *D*.[6]

Hattiangadi distinguishes between *discriminatory* and *common* problems. A problem is discriminatory between two (or more) traditions iff it belongs to just one tradition. A problem is common iff it belongs to all.[7] To represent discriminatory problems, Hattiangadi uses diagrams of the following kind (see Hattiangadi [1979], p. 58, schema 6):

Common and discriminatory problems

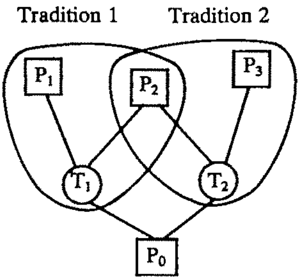


FIGURE 3

In Figure 3, there are two different traditions, TR_1 and TR_2. P_2 is a common problem, P_1 and P_3 are discriminatory. Now, assume that the debate in Figure 3 evolves as in Figure 4:

---

[5] There is a further problem with this definition. Hattiangadi states the definition for lines, but it is clear that they are main lines, so that they correspond to what I call here traditions (see also footnote 2).

[6] I will comment more on these evaluation-concepts in Section 3.4, since they are the basis of Hattiangadi's methodology (see also footnote 15 and Section 4). For the definition of value of a solution see Hattiangadi [1979], p. 60: 'The greater . . . the tradition'.

[7] There is a minor difficulty with this terminology. Hattiangadi also uses discriminatory in a different sense. A discriminatory problem is a problem whose eventual solution affects the standing of two traditions. If we take this meaning of discriminatory, then all non-common problems, and at least some common problems, are discriminatory, so that, if we are not careful in distiguishing the two senses, we arrive at a plain contradiction: some common problems are both discriminatory and not discriminatory. To avoid this unpleasant consequence, we can retain only the first sense of discriminatory (=non-common). The second sense can be captured by the concept of *value* of a solution.

The debate in fig. 3 at a later stage

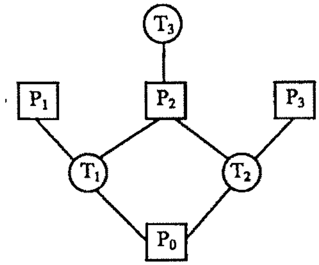


FIGURE 4

The problem with Figure 4 is that we do not know to which tradition T_3 belongs. Is T_3 a solution of P_2 in TR_1, a solution of P_2 in TR_2, or both? The representation of common problems in tree-like diagrams must be improved, if they are to be used to appraise competing traditions by means of the evaluation-concepts previously defined. This is formally accomplished if we represent a common problem as a divided box. Figure 4 will thus become one of the three diagrams in Figure 5.

Possible modifications of fig. 4

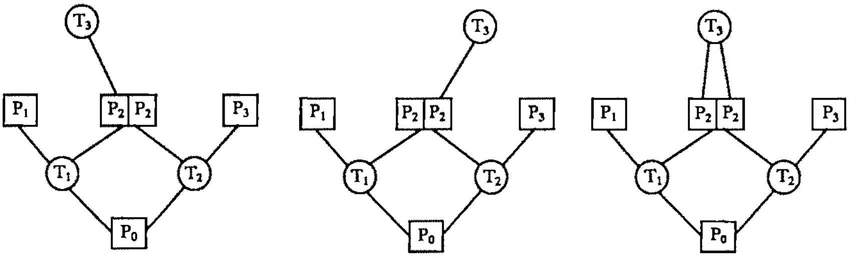


FIGURE 5

It must be noticed, however, that this graphical device has only solved the problem of correctly representing traditions (or lines) once they have been identified. A more substantive difficulty is still open: how could we actually choose one of the three possibilities of Figure 4? Hattiangadi's theory does not provide any answer to this question. One would expect the theories in the same tradition to be somehow related, for instance in a Lakatosian way—they all share the same basic assumptions or hard-core. Hattiangadi, however, does not introduce any notion of this kind, so that the identification of different traditions exclusively relies on the careful historical study of any specific case.

3 CRITICAL COMMENTS

3.1 *Criticism of Hattiangadi's Historical Theory of Problems*

Hattiangadi's historical theory suffers from a diffuse instability in the relations between its main concepts. So it poses, in the first place, a quite intriguing *formal problem*. Its central meaning is, however, rather definite, so that the interpretative problem can be solved. The interpretation given in Section 2 is a step in this direction, and it might be further elaborated, so as to arrive at a completely adequate formulation. The formal inadequacies of the theory are basically due to the ambiguous usage of the concept of *line*, which sometimes refers to traditions, and, in other occasions, refers to lines.[8] This ambiguity also leads to deficiencies in the formulation of the notions of dominance, standing, viability, and value of a solution. However, when the distinction between lines and traditions is introduced, and the concept of line is sufficiently elaborated, all these difficulties can be overcome.

But Hattiangadi's historical theory also has some *substantive problems*. These are concerned with the two properties of lines, conservation and non-recurrence (see footnote 3), which are essential for Hattiangadi's view of problems as historical entities. To modify (or to drop) these two properties would amount to completely changing his theory. But this is precisely what is required, for non-recurrence and conservation are not properties which can reasonably be attributed to real-life scientific problems.

According to the non-recurrence property, if a problem, *P*, is solved by some theory, in a line, *P* never occurs again as a problem of some later theory in that line. This property is too strong, for it a-priori rules out cases which can obviously occur. Consider the following example. For Aristotle's biological theory, the appearance of a mutation constitutes a problem, since a particular organism deviates from the essential characters of its species. On the other hand, mutations also are a problem for Darwin's theory of evolution, since his theory requires the occurrence of variations, but no mechanism is provided to account for such a phenomenon. We thus see that two different theories in the same line have the same problem, namely, to explain the phenomenon of mutations. Given Hattiangadi's theory, there are two possible replies to this objection. First, one could deny that Aristotle's essentialist theory and Darwin's theory were in the same line. Or, second, one might claim that although Darwin's problem is similar to the earlier one, it is not identical, because it possesses an additional historical structure. The difficulty with these two replies, however, is that Hattiangadi's concepts of line and problem are too vague to rule out the interpretation that Darwin's and Aristotle's theory were indeed in the same line and that they share the same problem.

Something similar happens with the conservation property: in a (closed) line any problem solved by theory *T* is also solved by any later theory in that line.

---

[8] See footnote 2.

This property is arbitrary. In fact it is possible to think of many cases in which two theories in a line do not satisfy such a property. It may be true that there is a continuous line between Parmenides's problem of motion and, for example, Newtonian mechanics. Nevertheless, it seems quite odd to affirm that Newtonian mechanics solves Parmedides's problem.

This arbitrariness of the two properties of lines would disappear if: (1) they followed from Hattiangadi's logical theory of problems; (2) such a theory were strongly justified by the history of science. But it will be shown below that neither (1), nor (2), is the case, so that Hattiangadi's historical theory turns out to be extremely implausible.

3.2 *Criticism of Hattiangadi's Logical Theory of Problems*

The main thesis of Hattiangadi's analysis of the logical structure of problems is that all intellectual problems are inconsistencies. I have already mentioned one strong motivation for holding this thesis. If all problems are inconsistencies, and if our seeking knowledge consists in trying to solve problems, then the answer to the question 'why do we seek knowledge?' is very simple and convincing: because we must for practical reasons.[9] The central role of inconsistencies for the development of science has also been stressed by Popper, since a refutation is nothing else than an inconsistency between a theory, hypothesis, or expectation, and a well corroborated basic statement. But, although many problems certainly are inconsistencies, still it seems obvious that there are problems which are not contradictions. At least two large categories of problems, gaps in our knowledge, and formal problems, are of this kind.

A typical knowledge-gap is the problem of measurement. We know that a certain object has a certain (physical) magnitude, but we do not yet know the value. There is no reasonable sense in which a problem of this kind might be interpreted as a contradiction. One might reply that this is not a deep, or scientific, problem. But it is well known that an important part of the scientific activity is concerned with just this kind of problem. For example, according to Newton's theory, the force between two unit masses at unit distance is the same for any two objects. But the value of this constant (the universal gravitation constant) was not known for about a century, until Cavendish devised an apparatus capable to determine such value.[10]

---

[9] It is, however, worth noticing that Hattiangadi's answer to this question heavily rests on the assumption that our 'natural' logic is classical. This assumption is not uncontroversial, it is for instance challenged by those logicians who believe that relevant, and not classical logic, is the underlying structure of our informal reasoning. Apart from this difficulty, Hattiangadi's answer also seems to overlook the effectiveness of pragmatic devices in keeping separate potentially inconsistent beliefs. It seems that our need of solving inconsistencies is not so urgent, and so frequent, as Hattiangadi claims.

[10] See Kuhn [1962], pp. 25–6: 'First is . . . sort of fact', and pp. 27–8: 'In the more mathematical . . . a stable solution'.

Formal problems concern the *formulation* of a theory. These difficulties may sometimes involve inconsistencies, but, if we reduce formal problems to inconsistencies, we lose sight of their peculiar character. Formal problems arise from unstable relations between the main terms of a theory. If we think of a theory as a net-like structure, then the nodes are its terms, and the lines are the statements in which these terms are involved. This structure may be unstable, in the sense that the theoretical statements may not be clearly formulated. The most stable part of a conceptual system is given by its terms (sometimes called concepts or categories), the most variable is given by the relations between terms. To any different configuration of the system, that is, to any fixed set of terms in some fixed relations, a meaning corresponds. If the system is not arbitrary, its meaning oscillates around a central point. The typical formal problem, relative to a system S, consists in trying to stabilize the system, while preserving the central part of its meaning oscillations. If we take this characterization of formal problems, we see that they are problems of interpretation, whose formulation does not essentially involve inconsistencies.

An interesting example of a conceptual system which has a formal problem is the present theory of natural selection. In such a theory there is no shared formulation of the principle of natural selection. However, there is a family of resembling principles (see Bradie and Gromko [1981]). The same holds for the concept of fitness. There are many different interpretations of this concept. Moreover, it we consider just one interpretation, for instance the propensity-interpretation of fitness, we discover that there is a family of similar definitions which employ the propensity-interpretation of probability in a very unprecise and questionable way.

The theory of the logical structure of problems also has a technical difficulty. For Hattiangadi, a *problem* is an inconsistent set of statements, I, together with the set, $S(I)$, of all the solutions of I (see Hattiangadi [1978], p. 357: 'A problem . . . the solutions').[11]

E is 'the' *explanatory power* of I[12] iff I is an inconsistent set of statements, and E is a consistent set of statements such that: [1] any of its members is a reasonable candidate for truth; [2] E is implied by a consistent subset of I; [3] for any consistent set of statements, E', which is not logically equivalent to E, and also satisfies [1] and [2], there is at least one statement, A, such that A is implied by E, and A is not implied by E' (see Hattiangadi [1978], p. 358: 'My fifth . . . for truth', and p. 359: 'Maximality condition . . . consequence of B'').

C is a *solution* of I iff I is an inconsistent set of statements, and C is a consistent set of statements which implies 'the' explanatory power, E, of I (see Hattiangadi [1978], p. 358: 'My fourth . . . set of statements').

---

[11] X is an inconsistent (consistent) set of statements iff $Cn(X)$ is inconsistent (consistent), where $Cn(X)$ is the set of all the logical consequences of X. Similarly for statements.

[12] Contrary to the usual terminology, according to Hattiangadi, the explanatory power of a theory, I, is not necessarily unique. This follows from his definition of this concept (see Hattiangadi [1978], p. 359: '*Corollary to* . . . selecting solutions').

Assume we have a problem, $P=(I, S(I))$, and let C be a solution of I, that is, C is member of $S(I)$. Since C is a *consistent* set of statements, there is no problem, P', such that $P'=(C, S(C))$. This leads to the following result: if the logical structure of problems is analyzed as above, then there can be no solution of a problem which, in turn, leads to another problem. Hattiangadi is of course aware of this difficulty. To overcome it, he allows two further possibilities:

(a) 'A problem may also be solved by an inconsistent set of statements, if the same problem does not arise again in the solution' (Hattiangadi [1978], p. 362);

(b) P is *essentially the same problem* as Q iff $P=(I_1, S(I_1))$, and $Q=(I_2, S(I_2))$, and the intersection of I_1 and I_2, I, is inconsistent, and $R=(I, S(I))$ is such that $S(I)$ includes both $S(I_1)$ and $S(I_2)$ (see Hattiangadi [1978], p. 362: 'Thesis seven . . . consistent subsets of X').

There are essentially two difficulties with theses (a) and (b). The first is as follows. Thesis (b) is a definition of a symmetric relation between problems. How should we interpret thesis (a)? If (a) is a definition of 'solution of a problem', when the solution is inconsistent, we have:

(a′) I_2 is an *inconsistent solution* of I_1 iff I_1 is an inconsistent set of statements, and I_2 is inconsistent, and Q is not essentially the same problem as P, where $P=(I_1, S(I_1))$, and $Q=(I_2, S(I_2))$.

From (a′), and since (b) is symmetric, we obtain that, for any two inconsistent sets of statements, I_1 and I_2, if I_2 is an inconsistent solution of I_1, then I_1 is an inconsistent solution of I_2. But it is quite obvious that the relation x is an inconsistent solution of y should not be symmetric. Consequently, (a) cannot be a definition of such a relation; (a), at most, gives a necessary condition.

The second difficulty regards thesis (b). Let us consider the example which Hattiangadi discusses on p. 362 of his [1978]. Let $X=\{A_1, \ldots, A_n, B, -B\}$, $Y=\{A_1 \ldots A_n, A_{n+1}, B, -B\}$, $P=(X, S(X))$, and $Q=(Y, S(Y))$. Assume that A_{n+1} is a consistent statement, and that it is not implied by any consistent subset of X. Hattiangadi affirms that Q is essentially the same problem as P. This claim is intuitively plausible, for Y and X have the same 'troublesome pair', B and $-B$, and Y differs from X just for A_{n+1}, which has been added. Nevertheless, this claim is contradicted by definition (b). In fact, the intersection of X and Y, I, is equal to X. Since A_{n+1} is consistent and it is not implied by any consistent subset of X, $S(I)$ does not include $S(Y)$. Consequently, by definition (b), Q is not essentially the same problem as P. If we take this example as an obvious case of two inconsistent sets which have the same problem (as Hattiangadi does), then definition (b) is not adequate.

In conclusion, Hattiangadi's analysis of the logical structure of problems, if not further elaborated, does not allow the possibility that an inconsistency of theory T_1 is solved by theory T_2, which is also inconsistent, but not 'in the same way'. Contrary to his claims, Hattiangadi has not given any explication of this concept, for theses (a) and (b) are not, respectively, adequate definitions of the relations x is an inconsistent solution of y, and x is essentially the same problem as y. If (a) and (b) are dropped, then Hattiangadi's logical theory implies that a solution of a problem can *never* lead to any other problem.

3.3 *The Relation Between the Logical and the Historical Theory*

Hattiangadi repeatedly claims that all problems share the same logical structure.[13] If this logical structure is the one discussed in par. 3.2, such a claim is unjustified. We have seen in Section 2 that one of the peculiar features of deep, and scientific, problems is that their solutions are theories which, in turn, have other problems. But we have also seen, in par. 3.2, that the logical explication of this idea is not tenable. Moreover, if the logical theory of problems is taken as it stands, then it is incompatible with such an idea. We can thus conclude that the logical theory of problems does not support the historical one. The two theories are also incompatible, unless the logical one is further developed in order to allow inconsistent solutions of inconsistencies.[14]

3.4 *Criticism of Hattiangadi's Methodology*

We have seen in Section 2 that the historical theory defines four evaluation-concepts:[15] domination, standing, viability, and value of a solution. Hattiangadi affirms that the definition of the value of a solution is the basic rule which permits us to understand every other rule regarding the evaluation of scientific theories (see Hattiangadi [1979], pp. 61–2: 'I suggest . . . by Popper'). This claim cannot be taken literally, for a definition is not a rule. The definition of an evaluation-term, however, can be the basis for a methodological rule, which, in this case, is:

(*R*) *given an intellectual debate, choose the solution with the highest value.*

Consider, now the intellectual debate in Figure 6.

---

[13] Besides the difficulties which I discuss in the text, this thesis is also affected by a minor one. Hattiangadi gives *two distinct* characterizations of the logical structure or problems. I have presented the first in the previous paragraph. The second is in Hattiangadi [1979], p. 50: 'Suppose . . . structure of problems'. It is never completely clear which logical structure (the first or the second?) all problems are supposed to share. I am assuming here that this common structure is given by the first theory, the 'official' logical theory, which Hattiangadi has discussed throughout his [1978] paper.

[14] It is not even clear whether the eventual solution of this difficulty would be sufficient for harmonizing logical and historical theory. The notion of a theory which has many, different, problems would also be needed. It is not prima-facie clear how an inconsistency-view of problems can accommodate this concept.

[15] What I call evaluation-concept is usually called standard, criterion, and, sometimes, (cognitive) value.

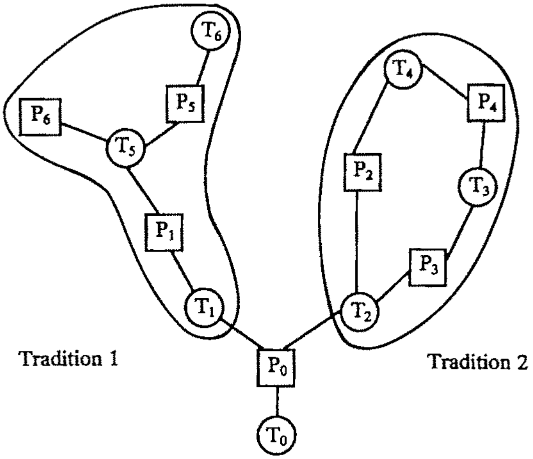


FIGURE 6

According to Hattiangadi's definition, the value of T_4 is higher than the value of T_6, since T_4 positively affects the standing of its tradition more than T_6 does. In fact, tradition 2, without T_4, has two open problems, while, with T_4, has no open problem. On the other hand, tradition 1 has two open problems in both cases. In accordance with (*R*), we should choose T_4. Hattiangadi maintains that:

(*a*) *choices based on (R) are rational;*
(*b*) *scientists really choose theories (and consequently traditions) in this way.*

Let us first notice that these two claims correspond to two different aspects of Hattiangadi's methodology. (*a*) amounts to saying that this methodology is a *normative* theory; (*b*) affirms that it is also *descriptive.*

Let us now consider claim (*a*), and the example in Figure 6. In what sense is choosing T_4, instead of T_6, rational? In the given debate, T_4 is better than T_6. But, at a later stage, the standing of the two traditions might be reversed. At that stage, it would be rational to choose the last theory of tradition 1. It then follows that choosing tradition 2 is rational at a certain stage and irrational at a later stage. This is a rather weak concept of rationality. Also, it is incompatible with the commonly accepted idea, which Hattiangadi seems to share, of a convergence-pattern in the development of science. According to Hattiangadi's model, tradition 1 and tradition 2 might be based on two incompatible, and very different conceptions of the world. It is possible to think of an intellectual debate in which the standing of the two traditions is regularly reversed at later stages. Our knowledge of the world would accordingly oscillate without showing any convergence.

There is also a paradoxical consequence of this view of rationality. Assume that thesis (*a*) is true, and that all the scientists choose theories according to rule (*R*), that is, they all are rational. Also assume that there are two competing traditions, TR_1 and TR_2, and that, very early in the debate, TR_2 comes to dominate TR_1. Then all the scientists of TR_1 will switch to TR_2. Assume that the standing beteen TR_1 and TR_2 remains favorable to TR_2 for a very long time (say 1,000 years), so that nobody, during this period, chooses and further develops TR_1. But now assume that, after 1,000 years, TR_2 starts accumulating problems, so that the standing between TR_1 and TR_2 gets reversed. At that moment, all the scientists will rationally choose the last theory of TR_1. In this way, however, 1,000 years of good scientific work and discoveries will be lost. This seems sufficient for concluding that claim (*a*) is not tenable.

Let us consider, now, claim (*b*). Whether (*b*) holds is, essentially, an empirical question. What 'choice-behavior' could confirm (*b*)? If (*b*) were true, then scientists should be particularly interested in the history, or genealogy, of the theories they choose. The reason is very simple. Let us consider Figure 7.

According to Hattiangadi's theory, T_4 is better than T_8, for T_4 makes TR_1 dominate TR_2. But, to realize that T_4 has a greater value than T_8, one must consider the whole intellectual debate. So, if scientists really applied Hattiangadi's rules, they first of all would carefully study the complete history of the theories they choose.[16] But this is not the case, therefore thesis (*b*) must be rejected.

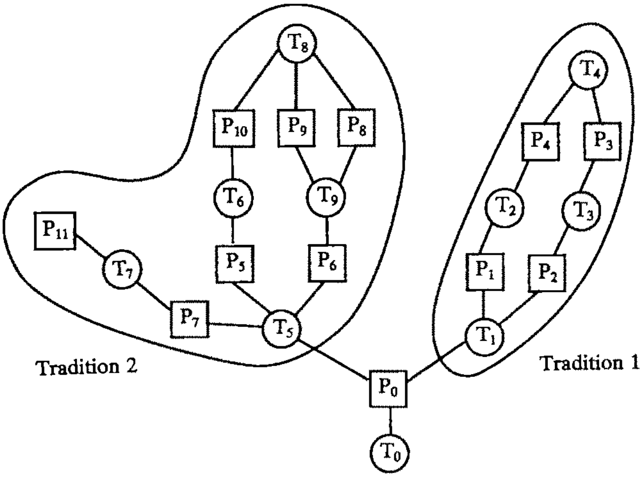


FIGURE 7

---

[16] If a scientist were not a good historian, he might easily choose the wrong theory. He might, for example, prefer T_8 to T_4, because T_8 unifies two theories, like T_4 does, but, in additon, T_8 solves more problems.

A possible reply to this objection is that, in fact, scientists have some knowledge of the historical background of the theories and problems they discuss, even though this knowledge is not acquired by means of a systematic study of the history of science:

> If we call certain problems which arise within intellectual disciplines 'deep' problems . . . then my thesis is that *the structure of 'deep' problems is historical in nature.*
>
> The intellectual who understands the structure of such a problem must, therefore, understand some distilled history of the problem. This is certainly not to say that all intellectuals are historians. In a similar case, a solicitor who knows Common Law may not be a historian of law, even though Common Law may not has a historical structure, insofar as its structure depends upon a history of practise and precedent. In a similar way the scientist studies problems that have a historical structure, yet with this difference, that the structure of deep problems is determined historically not by practise and precedent but in a different manner altogether. (Hattiangadi [1979], pp. 52–3)

What Hattiangadi affirms may well be true. Still, this is not sufficient for justifying thesis (*b*). In order to do this we should be able to show that the 'distilled history' that any scientist understands comprehends the whole traditions to which the given problems belong. In fact, if we do not make this assumption, we cannot maintain that scientists really choose theories according to Hattiangadi's rule (*R*), for this rule, together with the definition of value of a solution (see Section 2) implies that a comparison must be made between the whole traditions under consideration. But this assumption is not plausible, so that the scientists' knowledge of a 'distilled history' cannot provide any justification for thesis (*b*).

4 CONCLUDING REMARKS: HATTIANGADI'S THEORY AND THE STRUCTURE OF STANDARD EPISTEMOLOGIES

To fully appreciate the theoretical significance of Hattiangadi's theory, we must now consider its general structure. Surprisingly enough, this structure is shared by a large group of theories of science, which, to a more superficial analysis, might well seem to be worlds apart. Theories with such a structure will be called standard (or traditional) epistemologies.

Standard (or traditional) epistemologies assume a *theory-choice postulate*, which can be put in this form:

> (*TC*) *scientists choose theories following certain rules, which are based on some evaluation concepts, unless disturbing factors impede them.*

Some standard epistemologies also assume another postulate, which can be called the *rationality postulate*:

> (*RC*) *scientists' rule-based choices are rational.*

(*TC*) and (*RC*), together are equivalent to the thesis of the rationality of the context of appraisal. If a standard epistemology also assumes (*RC*), it is a *normative* epistemology.

The basic problem of all standard epistemologies consists in specifying the *rules* which, according to (*TC*), guide the scientists' choices. This problem leads to the problem of specifying the *evaluation-concepts* on which such rules are based, and this, in turn, leads to the problem of specifying the nature, or *structure of the theories* which the evaluation-concepts are supposed to evaluate.

Consequently, the conceptual architecture of any standard epistemology is as follows. In the first place, there is a *structural theory*,[17] that is, a theory of what a theory is. This structural theory supports definitions of *evaluation-concepts*,[18] which in turn support the formulation of *choice-rules*.[19] When such rules are specified, *a specific version*, (*TCS*), of the *theory-choice postulate*,[20] is assumed. In general, (*TC*) follows from (*TCS*),[21] therefore any specific epistemology with this structure is a model of a standard epistemology. Specific normative epistemologies also assume *a specific version*, (*RCS*), of (*RC*).[22] The *methodology* of a standard epistemology is the set of its rules together with (*TCS*) and, eventually, (*RCS*).[23]

The most original feature of Hattiangadi's epistemology consists in that the whole system is based on the concept of scientific problem. First, for Hattiangadi, the very notion of scientific theory involves a reference to problems. Theories are conceptual systems in which problems arise, and which are solutions of previous problems. Theories are thus essentially linked to problems, so that we cannot separate them either from the problems which they are designed to solve or from the problems which they raise. This is an important insight about the nature of scientific theories, an insight which has also been expressed by Popper:

> My fourth point is that every attempt (except the most trivial) to understand a

---

[17] Hattiangadi's structural theory is provided by his historical theory of problems. More precisely, a scientific theory is defined as the solution of a problem in a given tradition.

[18] Hattiangadi's evaluation-concepts are: dominance, standing, viability, and value of a solution (see Section 2).

[19] Hattiangadi's choice-rule is (*R*) (see Section 3.4).

[20] Hattiangadi's specific version of the theory-choice postulate is assumption (*b*) (see Section 3.4).

[21] In fact, in Hattiangadi's epistemology, (*TC*) follows from (*b*) (see Section 3.4).

[22] Hattiangadi's specific version of the rationality-postulate is assumption (*a*) (see Section 3.4).

[23] Instances of standard epistemologies are: logical empiricist views of science, Popper's, Lakatos's, and Laudan's epistemologies. All these theories are also normative. Examples of non-standard epistemologies are Kuhn's and Feyerabend's theories. These theories are non-standard because they assume the negation of the theory-choice postulate. In fact, according to Kuhn Feyerabend, there are no shared rules for the acceptance or rejection of theories. Toulmin's theory see [1972]) might be classified as 'semi-standard', since the theory-choice postulate is accepted in a weaker form—scientists choose theories according to rules, but these rules vary over time. The most recent developments of Laudan's thought (see [1984]) go in the same condition.

> theory is about to open up a historical investigation about this theory and its problem, which thus become part of the *object* of the investigation. (Popper [1972], p. 177)

Second, the evaluation of theories is ultimately based on the consideration of problems.[24] In fact, the four evaluation-concepts which Hattiangadi proposes (dominance, standing, viability, and value of a solution) all involve an essential reference to problems (see Section 1). This idea distinguishes Hattiangadi's view from the majority of the standard epistemologies, which are mostly concerned with concepts like (empirical) content, accuracy, simplicity, fruitfulness, confirmation, corroboration, *etc.* Since these concepts have not led thus far to a satisfactory theory of science, Hattiangadi's problem-based alternative deserves careful attention. Nevertheless, the critical analysis of the preceding section has also shown that Hattiangadi's proposal for a new epistemology has serious difficulties.

We can also wonder whether Hattiangadi has fully exploited the idea that focusing on problems is crucial for a correct understanding of the development of science. This doubt leads us to question the very structure of all standard epistemologies. These systems essentially focus on the process of theory-choice, which is accorded a sort of special status for understanding science. The underlying idea is that if we were able to explain how scientists choose theories, we would have the key for understanding the whole scientific activity and its historical development. Even though this assumption has been, and still is, widely accepted by philosophers of science, its implausibility is quite striking. The process of theory-choice is only a part of the whole complex of scientific activities, and to assume that the dynamics of science can be understood by primarily considering just this process is a conjecture which needs some justification. There is no prima facie evidence in favor of this assumption, but instead there is indirect evidence against it, for all standard epistemologies have thus far failed to give a satisfactory account of the overall development of science.

There is perhaps a historical reason for the central place which the study of the process of theory-choice has acquired in the contemporary philosophy of science. This can be traced back to Reichenbach's distinction between the context of discovery and the context of justification, and to the long accepted view that philosophy of science could only be concerned with justification, while discovery should be investigated by other disciplines, such as psychology, sociology, history of science, *etc.* This view is connected to the idea of an essential difference between justification and discovery. While the first would be a rational process, the second would lack this character. Therefore, only justification could undergo a philosophical analysis, while discovery would be the object of specific scientific disciplines. But the clearcut distinction between the two contexts is no longer tenable in the light of the most recent epistemological debate, so that also the central role of the theory-choice process needs to be revised.

---

[24] Laudan's book *Progress and Its Problems* ([1977]) also stresses the importance of problems for the evaluation of theories (see also footnote 1).

Laudan's analysis of the context of pursuit might seem a first step in this direction. For Laudan, to understand the dynamics of science, we must not only consider the theories which scientists accept, but also the theories on which they choose to work, without necessarily committing themselves to their acceptance. Laudan's analysis of pursuit, however, exclusively focuses on the conditions under which the *choice* to develop a theory is rational. Hence, also Laudan's epistemology maintains theory-choice as its primary object.

If philosophy of science aims to understand science, and its development, in all the relevant aspects, then, besides the context of justification, also discovery is its proper object. Perhaps the most interesting feature of a problem-based approach is that it seems to provide a framework for a unified account of both justification and discovery. Hattiangadi's epistemology shows that the concept of problem *can* be used for constructing an explanation of how theories are accepted or rejected. Hattiangadi, however, has not exploited the full potential of this approach, since he has not tried to explain in detail how problems lead to the development of existing theories and to the creation of new theories. An adequate theory of science should deal with these issues, and also with the more traditional questions about acceptance and rejection. Focusing on problems seems to be a promising strategy for elaborating such a theory.

Department of History and Philosophy of Science
Indiana University

REFERENCES

BRADIE, M. and GROMKO, M. [1981]: 'The Status of the Principle of Natural Selection', *Nature and System*, **3**, pp. 3–12.

HATTIANGADI, J. N. [1978]: 'The Structure of Problems' (part I), *Philosophy of the Social Sciences*, **8**, pp. 345–65.

HATTIANGADI, J. N. [1979]: 'The Structure of Problems' (part II), *Philosophy of the Social Sciences*, **9**, pp. 49–76.

KUHN, T. S. [1962]: *The Structure of the Scientific Revolutions*, Second Edition 1970. Chicago: The University of Chicago Press.

LAUDAN, L. [1977]: *Progress and its Problems*. Berkeley: University of California Press.

LAUDAN, L. [1984]: *Science and Values*. Berkeley: University of California Press.

POPPER, K. R. [1972]: 'On the Theory of The Objective Mind', in Popper K. R., *Objective Knowledge*, revised edition 1979, pp. 152–90. Oxford: Clarendon Press.

TOULMIN, S. E. [1972]: *Human Understanding*. Princeton: Princeton University Press.

Brit. J. Phil. Sci. 39 (1988), 441–448 *Printed in Great Britain*

The Darwinian Synthesis: A Critique of the Rosenberg/Williams Argument

G. VAN BALEN

Over the past twenty years several authors—in particular Rosenberg [1979], Grene [1981] and Williams [1970] in declining order of explicitness—have argued that there has not really been a synthesis between Darwin's theory of natural selection and Mendel's theory of heredity. In this paper I will show that their argumentation fails, because it is based on an incomplete reconstruction of Darwin's theory.

My critique of the so-called Rosenberg/Williams argument consists of two related parts. First I will argue that Rosenberg's reconstruction of Darwin's theory leaves out one important assumption inherent in Darwin's theory, *viz.* the assumption that natural selection operates on small, continuous, randomly produced variations. The justification of this assumption constituted a serious problem to Darwin's theory, which was only to be solved by the Mendelian theory of heredity.

Secondly, I will argue that the axioms from Williams' formalization of Darwin's theory, which also furnished the basis for Rosenberg's reconstruction, would have been accepted by many rival evolutionary theories of Darwin's day. Most evolutionary theories, including Darwin's theory, shared Williams' basic model; they were distinguished from each other by each of them making *additional* assumptions. These assumptions, like the one that Darwin made, all concerned the question as to which variations are inheritable and which are sensitive to selection. Taking these arguments together I will be able to conclude that Darwin's theory really needed a specific theory of heredity to back up the one important assumption by which it was distinguished from its rival evolutionary theories.

I

There has been no synthesis, according to Rosenberg, Williams and Grene, because the theory of natural selection is a self-sufficient theory which assumes heredity, but which has no need of any alliance with a particular theory of heredity. The Mendelian theory of heredity gave support to Darwin's theory, but any other theory of heredity would have done as well. The relation between both theories was rather one of sustenance than of synthesis (Rosenberg [1979]).

Darwin's theory, according to Rosenberg, is 'starkly simple' and based on easily accessible observations. Given the observations that organisms reproduce themselves in geometric proportions and yet the number of individuals in a population remains constant, added to the fact that there are great individual variations among the organisms, the following conclusion may be inferred:

> '. . . since any population can produce more young than the region they occupy can maintain, there will be a struggle for existence among the organisms produced; and, in this competition, those organisms will survive, whose variations best suit them to the immediate environment—that is, a process of natural selection takes place. Add to these conclusions the further hypothesis that those selected variations may be inherited, and the result is the heart of Darwin's theory.' (Rosenberg [1979], p. 5.)

If this is what Darwin's theory roughly amounts to, then Rosenberg's argument seems reasonable enough. As long as selected variations may be inherited evolutionary change takes place. What mechanism inheritance follows is independent of Darwin's theory. But Rosenberg seems to overlook that Darwin made one additional claim which is essential to the theory of natural selection, namely the claim that the mechanism of evolution is the natural selection of *small* inheritable differences. It is this claim—not the basic tenets mentioned by Rosenberg—that led to many of the controversies at the turn of the century. Moreover, most modern biologists agree that the appearance of a consensus concerning the validity of this claim should be considered as an essential component of the modern synthesis (*cf.* Provine [1980], p. 339). Darwin's claim that natural selection operates on small differences, became an essential part of our present evolutionary theory. This historical fact would be particularly relevant to the Rosenberg/Williams argument in case these authors were to object, that they did not discuss Darwin's actual theory but only 'the part of the present theory for which Darwin provided all the crucial explanatory principles' (Williams [1970], p. 344). It shows that Darwin's small variations-postulate *is* one, if not the most important, of the 'crucial explanatory principles' which Darwin contributed to our present theory.

If Darwin assumed that natural selection primarily operates on small differences, then we have to add to his theory the further hypothesis that *small* continuous variations may be inherited *and* selected.[1] This addition completely alters the situation. In Rosenberg's characterization of Darwin's theory, evolution is dependent on some form of inheritance. Selection will follow as a matter of course. In my characterization, evolution is dependent on the inheritance *and* selection of small variations. No biologist, even in Darwin's time, would have objected against some form of inheritance, but many biologists at the turn of the century raised doubts either about the heritability of small variations or about the selection of small variations. For instance, one of the aims of Bateson's research program was to show that small, continuous variations are not inheritable (*cf.* Van Balen [1986]). If he had succeeded, Darwin's theory of natural selection would have been refuted. So it seems, in spite of Rosenberg's conclusion, that Darwin's theory really needed a specific theory of inheritance which would account for the heritability and the stability (to avoid swamping) of small variations.

---

[1] As is well-known, Darwin did not succeed in escaping from the serious imputation that even if small variations are inherited, they would not be stable enough for selection. In a critique of Darwin's own theory of heredity, his doctrine of pangenesis, Fleeming Jenkin showed that every variation, advantageous or not, would be swamped out of existence in two or three generations. Darwin's theory of heredity appeared to be inconsistent with his theory of natural selection.

It should be noted here, that I am not claiming that Darwin's theory needed *Mendel's* theory to be completed; I am only claiming that Darwin's theory needed a *specific* theory like Mendel's theory. But this is sufficient to counter the Rosenberg/Williams argument.

An important philosophical issue underlying their argument is whether evolutionary theory is spatio-temporally unrestricted.[2] In their view Darwin's theory provides a body of spatio-temporally unrestricted laws which can be supplemented with what ever theory of heredity appropriate for application anywhere in the universe. In my reconstruction, however, there are possible worlds in which Darwin's theory—or that part of Darwin's theory which we use in the present theory—is not applicable. For instance, in a world in which only large type-variations are inheritable and in which small, continuous variations are merely environmentally induced fluctuations of these type-variations,[3] there would not be any variations for Darwinian selection to act upon. For Darwin's theory to be applicable it was (logically) necessary that small, continuous variations be inheritable, and consequently it was necessary that Darwin's theory be completed by a specific theory which guaranteed this inheritability.

To put it in a slightly different phrasing: the Rosenberg/Williams claim, that for Darwin's theory to be applied in any empirical situation (and thus to be tested) it is only necessary to supplement it with a theory of heredity, is insufficient. There are a priori constraints on the empirical situations to which Darwinian theory can be applied. Only if it is empirically true that small, continuous variations are inheritable, will Darwin's theory be applicable. For this reason Darwin needed a specific theory of heredity. It is history's irony that Mendel's theory fulfilled this need, although initially it was taken to refute Darwin's theory.

Thus far I have argued that Rosenberg's argument fails, because he omits, deliberately or not, the small variations-postulate which I showed to be an essential part of Darwin's theory. This omission becomes even more prominent in the next section where I will argue that Williams' axiomatization, which furnished the basis of Rosenberg's reconstruction, should be seen as a model not only of Darwin's theory but of most rival evolutionary theories of Darwin's time. Most evolutionary theories, such as the various forms of saltationism, Lamarckism and orthogenesis, shared this basic model, while they were distinguished by each of them making different additional assumptions, having the same function as Darwin's small variations-postulate.

---

[2] This was brought to my attention by Mary Williams.

[3] This was the world the saltationists thought they lived in. Ironically, this conception of the relation between small, continuous variations and large type-variations, which comes from Johannsen, constituted the first adumbration of our modern genotype-phenotype distinction.

First it will turn out to be not surprising that the Rosenberg/Williams reconstruction is incomplete—as a model of *Darwin's* theory—because it appears to be a model of several mutually incompatible evolutionary theories. But secondly, and more importantly, it will turn out that the rival evolutionary theories primarily differed with respect to their preliminary solutions of the problem as to which variations were to be considered as relevant for evolution. The fact that there existed rival solutions to this problem seriously increased Darwin's need for an adequate theory of heredity.

II

To support his thesis Rosenberg refers to Williams' axiomatization of the theory of natural selection. Williams, too, states that Darwin's theory does not depend on any particular theory of inheritance (Williams [1970], p. 380). Accordingly, Mendelian genetics is not mentioned in Williams' mathematical model.[4] Williams' model is prone to a critique that is similar to the critique with respect to Rosenberg's more informal characterization. Again, Darwin's claim that evolution proceeds by the natural selection of small variations appears to be unaccounted for. What's more, I suspect that the five axioms, as they are reconstructed by Williams from Darwin's theory, would have been accepted by any biologist in Darwin's time, whether he be a Darwinian, Lamarckian or saltationist.[5] Williams' model is not an axiomatization of *Darwin's* theory of evolution, but an axiomatization of evolutionary theory *pure and simple.*[6] All, or at least most, rival evolutionary theories, including Darwin's theory, seem to be different interpretations or rather extensions of the basic model.

In order to justify these contentions by means of examples, we will first have a look at the five axioms in Williams' non-formal 'translation' (Williams [1973]; for the formalized version see Williams [1970]):

---

[4] This deliberate omission has been criticized by, among others, Ruse [1973].

[5] Of course, this does not alter the fact that Darwin has to be credited for first formulating these axioms informally.

[6] The basic tenets, axiomatized or not, are perhaps so 'starkly simple', to quote Rosenberg, no biologist would ever dare to question them.

1. Every Darwinian subclan is a subclan of a clan in some biocosm.
2. There is an upper limit to the number of organisms in any generation of a Darwinian subclan.
3. For each organism, b(1), there is a positive real number, 0 (b1), which describes its fitness in its environment.
4. Consider a subcland D1 of D. If D1 is superior in fitness to the rest of D for sufficiently many generations (where how many is 'sufficiently many' is determined by how much superior D1 is and how large D1 is), then the proportion of D1 in D will increase during these generations.
5. In every generation *m* of a Darwinian subclan D which is not on the verge of extinction, there is a subcland D1 such that: D1 is superior to the rest of D for long enough to ensure that D1 will increase relative to D; and as long as D contains biological entities that are not in D1, D1 retains sufficient superiority to ensure further increases relative to D.

Some clarification is necessary, especially with respect to the terms clan, subclan, Darwinian subclan and subcland. A clan may be any set of organisms if it is a temporally extended set of related organisms. In extremis, a clan may be the set of all organisms.[7] A subclan is a sub-set of a clan. A Darwinian subclan is a subclan which acts as a unit with respect to selection. Roughly this means that, while a subclan may consist of different (*e.g.* isolated) parts which are subject to different selective forces, this is not the case with a Darwinian subclan.[8] The subcland, finally, is a proper sub-set of a Darwinian subclan which contains organisms with a hereditary trait that gives them a selective advantage over their contemporaries.

Williams does not put any restrictions to the size of the clan(s) and the subclans (except the subcland, which has to be a proper sub-set). Depending on the interpretation of the model, a subclan may be a collection of related species or a small isolated population within a species. This serves her well, because it gives Williams the opportunity to throw some light in the units-of-selection-controversy: different levels of selection are different models of the theory (Williams [1973]). But this laudable flexibility gives us also the very reason why Williams' axiomatization cannot be an axiomatization of Darwin's theory of natural selection. Even a thoroughbred non-Darwinian theory like the classical saltationary theory can be shown to be a model of Williams' theory. If the minimal size of the subcland is taken to be precisely the species-size,[9] then we get the saltationary theory of evolution. The saltationists never denied some role to selection, but they made the important restriction that selection has nothing to do with speciation. Species arise by some other mechanism, i.e. mutation. So, if some new species is taken to be the subcland in axiom 5, then this subcland will spread in the 'Darwinian' subclan (here *e.g.* the clade) if it is superior to the rest of the 'Darwinian' subclan (*e.g.* the other species in the clade). The new species will be selected. So axiom 5 appears to be compatible with saltationary theory. The same holds good for the other axioms. The saltationists would have felt comfortable with this set of axioms, though they would not value them highly, since it is speciation which is really interesting.

---

[7] Because of the required ancestor-descendent relation a clan and its subclans are *ordered* sets.

[8] Note that there is nothing necessarily *Darwinian* about a Darwinian subclan. *E.g.*, in a typological conception of species a species may be a subclan which acts as a unit with respect to selection. This subclan would not be 'Darwinian'.

[9] There is no need to show how one would specify either a minimal size for a subcland or a size for a species. The point is, that given the definition of a subcland, for the classical saltationist the subcland can never be a sub-set of a species. If he would tolerate that, he would automatically admit the function of selection in speciation. Thus the classical saltationist will accept the five axioms if the smallest subcland is the species.

With respect to other evolutionary theories, for example Lamarckian theory,[10] the case seems even more simple. All five axioms are at least compatible with this theory. The most important difference between the Lamarckian theory and Darwin's theory is that in the latter case adaptation is explained by selection, while in the former case selection is explained by adaptation (the variations are already adapted 'pour le besoin', so it is no wonder they are selected). There is nothing in the axioms which refers to this difference. For the Lamarckians selection is relatively unimportant, but this does not alter the fact that Williams' axioms might as well have been part of Lamarckian theory as of Darwin's theory. No evolutionary theory completely excluded selection; '. . . even the severest critics of Darwinism have usually given selection some role, albeit a minor one' (Ruse [1982], p. 143).

Williams' axiomatization leaves room for different interpretations, which lead to mutually incompatible evolutionary theories. It seems to give us the common characteristic of all genuine evolutionary theories. In itself, of course, this is an important achievement, but it fails to be useful in clarifying the issue we are dealing with. We are interested in the problem whether a particular instantiation of the model, *viz. Darwin's* theory of natural selection, as it is distinguished from other evolutionary theories—even in our present conception—has been synthesized with Mendel's theory or not. And this problem is deeply related to the question of how Darwin's theory managed to survive amidst its rival evolutionary theories, which all shared the same basic model. The answer to this question can be found by comparing the additional assumptions made by these different theories.

These assumptions should be seen as partly tentative answers to questions like: What variations contribute to fitness? How are variations originated? Which of these are inherited? Which selected? Different answers resulted in different evolutionary theories. Darwin assumed that the evolutionarily relevant variations are small and randomly produced.[11] In other evolutionary theories the evolutionarily relevant variations are produced by 'orthogenesis' or 'pour le besoin'. For the saltationists only large variations, constitutive of new species, are evolutionarily relevant. All these rival theories had no quarrel about the basic tenets of evolutionary theory, first formulated by Darwin and now axiomatized by Williams.[12] Most controversies were primarily concerned with the presumed nature of the variations relevant to evolution.

---

[10] Note that it is *Lamarckian* theory that I am referring to. Most Lamarckians developed one or another form of a theory of the inheritance of acquired characteristics, which was not the main component of Lamarck's own theory. I am not sure whether *his* theory would be compatible with the five axioms, given his emphasis on an almost inevitable progression in all forms of nature.

Should the Lamarckians have succeeded in showing that only those variations are selected which are produced 'pour le besoin', then we would have had a Lamarckian theory of evolution. Should the saltationists have succeeded in showing that only large variations are inherited and selected, then Darwin's theory would have been refuted and the classical saltationary theory would have survived. The choice among these theories primarily depended on the solution of the problem of variation and the larger part of this solution would have had to consist of an appropriate theory of heredity.

A theory which states that only a specific part of the variations, that are possible in organisms, are evolutionarily relevant, would profit if it was backed up by a suitable theory of inheritance for at least that part of the variations. This was no less the case with Darwin's theory. In our world it was Darwin's theory that survived, because it was to be coupled with a theory of heredity which definitely demonstrated the evolutionary relevance of small, continuous, random variations. This coupling not only strengthened Darwin's theory, but also completed it to a fully adequate evolutionary theory. So I believe to have shown, *contra* Rosenberg, Williams and Grene, that a synthesis took place in the 1920s, at least in the sense that Darwin's theory acquired a *necessary* supplement in the form of Mendel's theory of heredity.

Philosophy Department
University of Nijmegen

REFERENCES

Grene, M. [1981]: 'Changing concepts of Darwinian evolution', *Monist*, 64, 195–213.

Provine, W. B. [1980]: 'Interpretative issues in the evolutionary synthesis—epilogue—', *The Evolutionary Synthesis*, Mayr, E. and Provine, W. B. (eds.). Cambridge, Mass., Harvard University Press.

Rosenberg, A. [1979]: 'Genetics and the theory of natural selection: Synthesis or Sustinance?', *Nature and System*, 1, 3–15.

Ruse, M. [1973]: *Philosophy of Biology*. London, Hutchinson University Library.

Ruse, M. [1982]: *Darwinism Defended*. Reading, Mass., Addison-Wesley Publishing Company.

---

[11] Later, forced by serious criticism, Darwin dropped the latter qualification and admitted the influence of 'use and abuse'.

[12] Not only these basic tenets were not quarrelled about, but in a fully explicit statement of the theories they would appear as *true and fundamental* principles for most of the theories.

Brit. J. Phil. Sci. 39 (1988), 449–467 *Printed in Great Britain*

Dimensions of Observability*

PETER KOSSO

ABSTRACT

The concept of observability of entities in physical science is typically analysed in terms of the nature and significance of a dichotomy between observables and unobservables. In the present work, however, this categorization is resisted and observability is analysed in a descriptive way in terms of the information which one can receive through interaction with objects in the world. The account of interaction and the transfer of information is done using applicable scientific theories. In this way, the question of observability of scientific entities is put to science itself. The result is a demonstration that observability has many dimensions, some more epistemically significant than others.

1 *Introduction*
2 *The Interaction-Information Account*
3 *Dimensions*
4 *Case Studies*
5 *The Declaration of Independence*
6 *Inference to the Source of Information*
7 *Summary*

I INTRODUCTION

Usually, discussions of observability of scientific entities are motivated by an urge to categorize and a plan to portray the proper, that is epistemically significant, application of the concept of being observable. Grover Maxwell argues [1962] that all entities are, in an important sense, possibly observable and hence observability is a poor measure of ontological status. This is motivated by the realization that there is no clear, nonarbitrary dichotomy between observables and unobservables. Bas van Fraassen responds that even without a sharp dichotomy there are some important differences in the observabilty of scientific entities, differences which, at least in clear cases of observables and unobservables, indicate the appropriate attitude toward belief in their existence. Thus van Fraassen ([1980], p. 16) offers a scheme to distinguish observables from unobservables. Dudley Shapere is also out to

---

* I am grateful to Philip Kitcher for his helpful suggestions and encouragement during the preparation of this paper.

categorize. In a careful execution of the plan to allow science to inform philosophy, Shapere presents 'a natural extension of what is to count as observational' ([1982], p. 506).

The approach of this paper is not to seek the proper application of the concept of being observable. In fact the concept of observability is politely eased off the stage, it being too suggestive of a dichotomy or, in its most liberal interpretation, of a one-dimensional continuum. Instead of observability we will ask about the general process of acquisition of information from the world and about the world. The plan is to point out the variety of ways that information-acquisition events can differ. What results is a method for studying observability, or rather, information acquisition. The method is then applied to a variety of examples and these case studies are used to point out the relative epistemic importance of the different degrees of freedom, the so-called dimensions of observability. The outcome of this exposure of the dimensions is to relocate the emphasis of a debate about scientific realism from observability (what is our warrant for belief in unobservables?) to a more fine grained concept, namely, one of the dimensions of observability. The point will be that observability is too complex (multi-dimensional) a concept to service the issue of realism and that the debate can be clarified and more productively pursued by focusing on one particular aspect of observability.

This approach shares with those of Shapere and van Fraassen the belief that science itself is the most appropriate adjudicator of observability, or the reliability of information acquisition. The method to be adopted is then one which invokes reliable scientific theories to describe the flow of information.

2 THE INTERACTION-INFORMATION ACCOUNT

The focus is on observation in physical science, understood in the general sense of acquisition of information. Science is in the business of gaining knowledge about the physical world, and observation, as it is part of science, must contribute to the knowledge. Observation must be not only a physical event but an epistemic event as well.

As an initial generalization then, one can say that observation is a manner of getting information of the physical world, from the physical world. It is accomplished with the conveyance of information from the world to the scientist, information that x is P, where x and P are features of the physical world. Conveyance of information is crucial and a simple event of interaction without information is not interesting insofar as epistemic processes are the topic. Events which might be described as observing x, for example, observing the table, are not epistemically valuable until we know what it is about x that is observed. We might grant as unproblematic that one observes, receives information, that the table is present or that the table is round. But our confidence decreases in the case of observing that the table is composed of molecules. The point is that the important epistemic value of an observation report is in its informational content, and informational content must include not only a report of the object x, but also a report of some property P, the what-it-is-about-x that is observed.

This motivates an application of observability as a two-place predicate.[1] This is because the nature of information specifies the two-place form, that —— is ——. To respect the information aspects of observation and observability it will be necessary to assess the observability of an object and a property together, rather than an object or property alone. Furthermore, an appeal to science as adjudicator of observability suggests that the description of the conveyance of information ought to be in terms of physical interaction. Interactions, as described by physics, psychology, and other sciences in between, are the way things get done in the world.

In accordance with these suggestions, a general characterization can be presented as a guide for studying cases of observability, that is, of information acquisition. It should leave open the possibility that observability may be a question of degree rather than simply a dichotomy:

> The ordered pair <object x, property P> is observable to the extent that there *can* be an *interaction* (or chain of interactions) between x and an *observing apparatus* such that the *information* that x is P is transmitted to the apparatus and eventually to a human scientist.

I have underlined the troublesome aspects of this characterization and insofar as ambiguous words like 'can' and 'observing apparatus' are left unanalysed, this description means very little. The crucial points in observability, after all, depend largely on the nature of the modality and the machinery used in the event. I will remedy these ambiguities by explaining in detail what is meant by the underlined terms.

The 'can', the modality of observability, and the 'interaction' are both concepts which can be handled entirely by the physical theories of the situation, the physics, biology, and other sciences of the source or medium. This is because physics and the specific physical laws which describe the object and the observer are exactly descriptions of the possibilities and nature of physical interactions. The fundamental laws of physics, for example, include specification of the circumstances under which elementary interactions occur. Electromagnetic interaction *can* happen when two bodies are electrically charged but not otherwise. The fundamentals of physics also include description of what happens in an interaction, and it is this kind of account of interaction which must be used to describe observability if that description is in the language of science.

---

[1] More generally one should apply observability as a many-place predicate to facilitate information of the form, 'that x bears relation R to y,' and 'that x bears relation R' to y and z, and so on. My specialized treatment of the two-place form can easily be generated to many-place forms.

Whether all information exchange can be accounted for by the physics of elementary forces is a question of the reduction of other sciences, and even the rest of physics, to the physics of elementary forces. It is wise to avoid this issue and allow as interactions not only the four elementary interactions but also groups of interactions which are manifested through other domains of science, as well as chains of interactions as mentioned in the above characterization.

The important point is that if one is to let science report on the nature of observability and observation, then observation must involve interaction. Interaction between x and an observing apparatus is a necessary condition for observation, that is, for acquiring information, but alone it is not sufficient. More conditions will have to be added, conditions relating to the nature of information.

But first consider another problematic concept in the characterization of observability, that of the observing apparatus. It is in the interest of open-mindedness that I consider the use of machines in observation. This account is to be a manner of studying the informational aspects of observability reports and it would be rash to ignore cases of information routed through machines. It is not the proper use of the term 'observable' which is the goal here. It is instead to develop and implement a method for describing the conveyance of information from the object to an observer, and then for evaluating claims of having received information from the object and about the object. In this characterization then I should initially allow the widest possible interpretation of the observing apparatus (or informing apparatus). It can be a human being, or it can be any apparatus which is capable of processing information from its own useful form, the form to which it reacts, into a form which is epistemically useful to humans. A simple thermometer is a good example. It processes information in the form of average kinetic energy of molecules into information in the form of the length of a column of mercury. The key is that the information which goes in must come out in a form which is accessable to the human scientist. The criterion of observability of <x,P> depends on whether such informational form is physically possible.

Clearly a lot hangs on the notion of information, and it is only by clarifying this that a full understanding of the observing apparatus is possible and that the necessary condition of interaction is augmented with sufficient, epistemic conditions. What is required is an explanation of what it is to have the information that x is P and how that information is conveyed in interaction. To this end I borrow heavily from Dretske's *Knowledge and the Flow of Information* [1981], and adapt his semantic theory of information to the purposes of observability in physical science. The motivation is simply to preclude disagreement over what it is that one observes. If we disagree over whether one can observe that the table is composed of molecules, then we will approach a resolution to this disagreement by invoking a shared notion of information. What information do we get by looking at the table?

Information content of a state of affairs can be specified quite precisely using a modified version of Dretske's definition of 'informational content' ([1981], p. 65).

> A state S has the *new information* that x is P if and only if $S, k \vdash (x \text{ is } P)$, but it is not the case that $k \vdash (x \text{ is } P)$.

k represents what is already known about the source x. For example, if in playing the game of peanut-under-the-shell we are initially told that the nut is under one of three shells, A, B or C, then the situation in which shells A and B are down and shell C is lifted with no peanut revealed contains the information that the peanut is under either A or B. This follows simply from the facts that (x is not under C) and (x is under A or B or C) imply that (x is under either A or B), but (x is under A or B or C) does not.

If one already knows that x is P, that is if $(x \text{ is } P) \in k$, then S cannot convey *new* information that x is P. But for purposes of observation one wants a concept of information that allows observation of what is already known. For this reason I augment the description of new information with one of redundant information:

> If $(x \text{ is } P\lambda) \in k$, then a state S has the *redundant information* that x is P if and only if $S, k \sim (x \text{ is } P) \vdash (x \text{ is } P)$, but it is not the case that $k \sim (x \text{ is } P) \vdash (x \text{ is } P)$.

Here the expression $k \sim (x \text{ is } P)$ means the body of knowledge k with the fact that x is P removed. I do not claim to be able to specify how this removal is done. But it is not essential in understanding what redundant information is to specify precisely how to recognize it. Redundant information is nothing more nor less than what would be new information if you did not know it already. In the characterization of observability the relevant concept of information is either new or redundant information.[2]

Information is transferred between states through interaction. The object in state S which has informational content (x is P) interacts with something else, the observing apparatus or some intermediate informational medium, with the result that this latter object is left in a state A which now has the information (x is P) whereas it did not have the information before the interaction. The information (x is P) may be nested (this is Dretske's term) in other information, such as (y is Q), of the A-state. Information (x is P) is nomically nested in the information (y is Q) just in case there is some law of nature which says that if y is Q then x is P. This is important in following the conveyance of information along a chain of interactions since the form of the information usually changes in interaction and it is only by unpacking what is nested that information of the object of observation is revealed. In a photoelectric interaction, for example, the initial photon state may contain the information that the photon has energy = 1 keV (x is P) while the subsequent electron state has the information that the electron has energy = ·997 keV (y is Q). Licensed by the physical law describing photo emission of electrons, and a knowledge of the binding energy of the material, the photon state information is nomically nested in the electron state information. The electron carries information about the photon.

---

[2] This description of information is exactly Dretske's except for the following modifications. What he calls 'information' I call 'new information' and then I add the description of redundant information. In place of his 'signal r' I use 'state S', a rewording he sanctions ([1981], p. 65). Where I describe information in terms of implication $\vdash$, he uses probabilities so that $S, k \vdash (x \text{ is } P)$ he writes as Prob $(x \text{ is } P|S,k) = 1$. My translation, I maintain, captures Dretske's meaning and presents informational content in a way which is more amenable to application to physical laws and less prone to misunderstanding. Dealing with logical implication rather than probabilities delivers us from difficulties in interpreting the probability theory.

Typically, actual cases of observing involve longer chains of interaction than the one-interaction example above. In these cases the information (x is P) of the object may get passed along the chain becoming nested more and more deeply with each interaction. The trick to observing <x,P> in this sense, and the focus of evaluating examples of observability claims, is to correlate, using relevant physical theories, the final apparatus-state information to initial object-state information.

The basics of the interaction-information account of observability are this: Observability, as an indication of information acquisition, requires that there can be an interaction chain from the object to a human observer or the final state of some appropriate observing apparatus. The final apparatus-state can be correlated to the object-state according to nomic relationships between the two, and in this way the observing apparatus has information about the object-state, information which comes from the object.

3 DIMENSIONS

This characterization facilitates an evaluation of the observability of <x,P> in terms of several degrees of freedom. The concept of observability is not one dimensional, and any careful account of observability must operate on all the dimensions.

One of those degrees of freedom, the one most likely to be thought of as *the* measure of observability is the *immediacy* of the S-state to a human observer. Roughly by this I mean an answer to the question, 'does it take some fancy machine to see this thing, or can I see it with my own (unaided) eyes?' Less roughly, by immediacy I mean a description of an object's potential to interact in an informationally correlated way. The more precise question is, 'with what can x interact in an informative way?' There are three kinds of answers to this question, and I use them as indicators of three possible values of immediacy.

One possible answer is that the object x simply cannot interact with any observing apparatus, whether it is human or otherwise, in an informationally correlated way. I will call such things *unobservable in principle.* It is a principled unobservability in the sense that the theories and physical principles which describe such things explicitly preclude their interacting in a way that allows information of the form 'that x is P' to be conveyed to an observing apparatus, any observing apparatus.

A second kind of response to the immediacy question is that the object in question can interact in a correlated way with some nonhuman observing apparatus but cannot so interact with a human sense organ. Call such an object *unperceivable in fact.* This terminology is potentially misleading and I should explain the motivation behind its use. An electron is a good example of something that is unperceivable in fact because it can interact with an observing apparatus, namely a bubble chamber, in an informative way, but cannot so interact with the human observer. But it is a contingent property of the world that the human body does not have bubble chambers as eyes and that it is not sensitive to electron information in some other way. It is a fact of the situation that the electron can be observed with the aid of machines, and a fact of the perceptual abilities of the body which make such things unobservable to the unassisted body.

The third value of immediacy is of those things which are *perceivable.* These are things which can interact in an informational way with a human being (or, allowing at most one, nonartifactual intermediary). This category is meant to cover those things that one can see, hear, feel, etc., without the assistance of observing tools. But even something so simply observable as the moon does not itself interact with the human body (well, it does gravitationally, but we are talking about seeing the moon), and hence the allowance for one intermediate interaction, a direct informational messenger. In the case of the moon, the medium is the group of photons which first interact with the moon and then with the pigment in the eye.

These three categories are meant to reflect an intuitive implementation of the interaction-information picture of observability. The unobservable in principle, unperceivable in fact, and perceivable descriptions represent, respectively, cases where no informational interaction is possible, where informational interaction is possible but only through observing tools, and where informational interaction is possible without tools.

The dimension of immediacy is just a first step in analysing the acquisition of information, an effort in *triage* to initially sort cases in terms of how best to treat them. The three values of immediacy are by no means detailed enough to separate importantly different cases. To add detail we must look at other dimensions, particularly of the middle class, the unperceivable in fact.

A second dimension which reveals the fine structure within the class of the unperceivable in fact is that of *directness.* This dimension is a modification of Shapere's direct/indirect dichotomy and it asks of the number of intermediate messengers that are required to convey the information from the initial object-state to the final observing-state, and the complexity of their physical journey. Cast in thoroughly information theoretic terms, the measure of the directness of observability is the measure of the number of interactions which would have to take place along the interaction chain from the initial-state to the final-state in order for the information to be conveyed. Simply put, how long must be the interaction chain from object to observer?

For example, information acquired with the aid of a magnifying glass is somewhat less direct than information obtained with the unaided eye, and information processed through a compound microscope is even less direct. This is determined by noting, in the first case, the extra interaction between photons and lens, and in the third case, the many interactions between photons and lenses. Each of these interactions is, in the physical science description of the process, a synapse in the conduit of information and its reliability must be accountable.

A third dimension of observability is the *amount of interpretation* which is necessary to correlate the final apparatus-state information to object-state information. This is intended to evaluate the nesting of object-state information in the final apparatus-state information. Given the apparatus-state information (y is Q), the amount of interpretation is a measure of how deeply nested is the object-state information (x is P). The measure can be quantified simply in terms of how many distinct physical laws are needed to get from (y is Q) to (x is P).

This dimension is different from the previous one, directness. Something which is observable via many interactions is not necessarily dependent on commensurately many laws to account for the information. Take Shapere's example of observing the interior of the sun using photons. This observation is hugely indirect since there is a very large number of interactions (and a large number of messengers) required to pass the information along. But since the interactions are nearly all of the same nature, the transmission of information is accounted for by a single physical law, and such an observation measures relatively low in terms of amount of interpretation. Amount of interpretation is a measure of epistemic closeness between the apparatus- and object-states. Directness, on the other hand, is a measure of the physical closeness between the two states.

A fourth dimension of observability can be identified, that of *independence of interpretation*. Given that there is a theory T of the object x (a theory which the observation of <x,P> is likely intended to test), one can ask whether the physical laws which account for the delivery of information in observation are independent of T. The answer is the measure of independence of interpretation. The manner and amount to which the theory T figures in the inference in the observation of <x,P> is closely related to the question of the theory dependence of observation as raised by Fodor [1984] and Shapere [1982]. It is a question of whether T can be isolated from the process of observing x, whether, in other words, only theories which are not theories of the object itself can suffice to support the observation. By 'support the observation' is meant that a theory is used to show that the information (x is P) is nomically nested in the final, observation apparatus information.

It is likely that the answer to this question will not be simply yes or no but will indicate a degree of independence. That this is the case will be shown in the following examples. In fact, the whole question of independence of interpretation is one which invites empirical answer. It is only by looking at the particular case and by analysing the conveyance and nesting of information that one can evaluate the degree of dependence of the observation on the theory T. We should be receptive to the possibility that T can be bifurcated into subtheories T1 and T2, where only T2 functions in observability accounting but T1 is responsible for the explicit description of x and P. In such a case there may be a degree of nepotism between T1 and T2, but not straightforward self-accounting by T1.

One can see *a priori* that there is a difference between the amount of interpretation and the independence or dependence of interpretation. The informational accounting of an observation could involve extensive inference, the information could be deeply nested, yet all of the inference could be independent in the sense that none of the laws accounting for the nomic nesting are T-laws. Or an observational process could involve only a single inference which involves T. The point is that there is no obvious connection between amount and independence of inference. The dimension of independence of inference does not measure epistemic closeness but epistemic nepotism.

The reason for pointing out all these dimensions of observability (and there may even be more) is to demonstrate the complexity of the concept. Classification of observability is neither a dichotomy between the observable and the unobservable, nor even a continuous spectrum along one axis. It is an evaluation of several parameters, locating the object-property pair in a many dimensional observability space. Evaluating each parameter requires consulting the physical science descriptions of the object x, the interaction-information chain, and the observing apparatus A-state. This interaction-information account is a guideline for using science to study observability.

The question of whether or not an object or even an object-property pair is observable is a bad question. It is a bad question because it ignores the complexity of observability. It looks for a one dimensional, discrete observability space where, in fact, that space is many dimensional and likely continuous. The more productive question is to ask instead about the observability characteristic of the object-property pair. Specifically ask, what is the nature of the information of the object-property pair and how does it get from the object to the observing apparatus. Then answer these questions by locating the example in observability space, that is by evaluating its components along the dimensions of observability.

Applying this account of observability acknowledges the complexity of the evaluation of observability of scientific entities, but it only begins the appreciation of epistemic significance. What is important to know now is the epistemic significance of each of the dimensions of observability. But before addressing that general question, consider some examples of observability and see the interaction-information account in action.

4 CASE STUDIES

Consider the observability of a single quark, or more appropriate to the object-property pair format, the observability of <quark, color>. This is a case of principled unobservability. The principle which prescribes the unobservability is color confinement which insists that quarks must occur only in colorless combinations, as color-anticolor meson states, or as red-green-blue colorless baryon states. The color of individual constituant quarks is never revealed in interaction with other particles or apparatus.

To say that color cannot be seen is just a sloppy way to claim that a colored particle cannot interact with another object which we regard as a part of the observing apparatus in such a way that the apparatus is left in a different state according to whether the interaction was with a red, green, or blue particle. The color of quarks is unobservable because physical interactions are color blind. It is not a problem with our eyes or with our machines. The unobservability is dictated by the physical laws of interaction themselves.

Even though single quarks are unobservable with respect to color, there are some other properties of single quarks which can interact with the rest of the world. Such properties are then possibly only unperceivable in fact and not unobservable in principle. Quark flavor is a property that can interact in a correlated way with an apparatus. Consider, for example, charm. It is not that individual charmed quarks can exist in isolation, but rather that a charmed quark can exist in a bound state with an uncharmed quark (as opposed to an anticharmed quark) such as an up-quark. The resulting meson $D^\circ = (c, \bar{u})$ is itself charmed since there is no anticharm to cancel the charmed quark. The charm of the quark shows through the meson in a way that color cannot. The charmed D° meson has a characteristic decay process indicating that the world can tell whether it is charmed or not. Charm, in other words, has an effect on interactions in a way that allows the <quark, charm> pair to be called only unperceivable in fact but not in principle. It is not charm which makes quarks observationally distinct from other elementary particles like neutrinos; it is color.

More will be learned about observability and the interaction-information account by considering another example, that of objects imaged with an electron microscope. The approach of the interaction-information account is to note the interaction between electrons and the specimen, to follow the electrons through the stages of the microscope, and to follow the transfer of information from specimen to electron-state and eventually to the human scientist. Observation, by this account, requires the reception of information *from* and *of* the specimen. The observability of the specimen is described by evaluating the dimensions described above. First I will briefly describe two of the many ways an electron microscope can be set up to exploit the interactive features of electrons to transmit information from microscopic objects.

For microbiological purposes one usually uses a conventional transmission electron microscope (CTEM). With an incoherent source, a CTEM works in the following way. A source of electrons is created by heating a piece of metal, that is, by thermionic emission. These electrons are accelerated in an electrostatic field and are focused into a narrow beam passing through a magnetic solenoid, usually referred to as the condensing lens. Typical electron energies are 100 keV to 1 MeV. The beam of electrons hits the specimen where some of the electrons are elastically scattered by atomic nuclei. Many of the scattered electrons are deflected by large angles from the uninterrupted beam direction, such that a small aperture placed just down-beam of the specimen will block these scattered electrons and allow only undeflected or slightly deflected electrons to pass. The electrons which pass through the aperture are then magnetically focused such that electrons from a particular point of the specimen are focused to a single point at the image plane.

At the image plane the electrons strike a phosphorescent screen or a photo emulsion plate creating a contrast field of light and dark areas corresponding to more or less electrons arriving. The features of this contrast field are informative of the features in the specimen since more dense areas of the specimen will scatter more incident electrons, sending them away from the beam to be blocked by the aperture. Thus fewer electrons will come through the aperture from the denser areas and these areas will be focused on the image plane to appear darker than the less dense, less efficient electron scattering areas. The relative densities of parts of an organic specimen, a cell, say, are usually enhanced by differential staining with a liquid of heavy atoms. The features of the specimen which absorb the stain will scatter more electrons and appear darker than those features which do not take the stain.

The wave qualities of the electrons can be utilized if the incident beam is coherent. This requires that the electron source produce electron waves which are at the same phase, and this is usually done by field emission, literally pulling electrons from a fine metal point with a strong electrostatic field. These phase-coherent electrons are targeted on the specimen as before and those that are scattered are phase-shifted by an amount δ which depends on the scattering potential of the target. (Particle scattering always produces a change of phase). The scattered electrons also have a longer distance Δd to travel to the image plane than do the unscattered electrons. The objective lens is tuned to reunite scattered and unscattered beams from each point in the specimen, producing an image point with an intensity which depends on the amount of constructive or destructive interference between phase shifted ($\delta + \Delta d/\lambda$) electrons and the unshifted beam. Accounting for the complicated interplay of lens abberation, lens focusing point, diffraction and interference between the many scattering centers, the contrast field produced at the image plane reproduces the scattering potential field of the specimen. This use of the electron microscope is known as phase contrast imaging.

The CTEM uses an electron beam which is wide enough to cover the entire area of interest in the specimen at once, just as an optical microscope illuminates the entire specimen being viewed. A scanning transmission electron microscope (STEM), on the other hand, operates with a finely focused electron beam (cross sectional area on the order of ·1 to 1·0 nm) which is incident on only small spots of the specimen at a time. The beam sweeps back and forth across the specimen and at each point the electron transmission is detected. The STEM functions much as the CTEM except that the image does not appear by phosphorescence at the image plane but by signals from an electron detector and the beam position, correlated and displayed on a television screen. As it is most typically used, the STEM differs from the CTEM in that it uses dark field imaging. Instead of detecting the electrons which pass through the circular aperture, a dark field image is produced by detecting the electrons which do *not* pass through. It counts those electrons which are widely scattered. This is done with an annular-shaped detector which reports many electrons when the incident beam is hitting a dense scattering center, and few electrons when the beam is not so heavily scattered. By ignoring the undeviated central beam the dark field imaging does not get information of phase difference between scattered and unscattered electrons.

Having outlined the basics of electron microscopy, we can consider two specific examples of things seen through electron microscopes. Consider first a CTEM, phase contrast image of a DNA molecule. Specifically, take the case of being informed by a CTEM image that the molecule is a closed loop. What one sees in this case is a simple pattern of dark lines on a lighter background. The reliability of the information, the assurance that this image corresponds to the shape of the specimen, is founded on an understanding of electromagnetism (to account for the emission of electrons and the effects of the lenses), the quantum mechanics of electron scattering, the chemistry of the stain used, the optics of diffraction, and the fluorescense of the screen. These are the physical theories which sanction the claim that the DNA is a closed loop.

Much of this theoretical support is likely to be forgotten if what you are interested in is DNA. This is because the information is displayed by the CTEM in an easily accessible way such that even someone unfamiliar with the mechanics of electron microscopy can quickly notice that the molecule is a closed loop. Contrast this with some biochemical test which, by some phenomenon of micro-organisms, indicates the same thing, namely that their DNA molecules are closed loops. Lehninger ([1970], p. 650) describes such a procedure which indicates that the DNA molecule in *E. coli* is structured as an endless, closed loop. The information is obtained in experiments by interrupting the poor *E. coli* at various stages of conjugation and noting the type and number of phenotypic traits in the progeny. In this case a glance at the results is not enough to lead one to the belief that the DNA is a closed loop unless one has a good idea of the whole scientific account which links a closed loop of DNA with the phenotypic manifestations. There is more theoretical preparation necessary to noticing that the strand is a closed loop in the biochemical test than with the electron microscope screen.

It seems as if the electron microscope is more a case of observing the DNA, or a more reliable information-acquisition event, than is the biochemical test, but it would be misleading to draw this distinction on the basis of the microscope presenting the DNA in a more readily noticeable form. It would be misleading because the biochemical results could conceivably be made to cause a television image of the closed loop which is theorized to account for the phenotypic results. That is, some form of transducer could be used to relay phenotype information, along with the nested information of the closed loop, into a machine which produces a picture of a closed DNA loop. The information of closed-loop DNA still arrives at the screen and is now easily noticed. Noticeability cannot sanction distinctions of observability in these cases. Rather it is the manner in which supporting theories are invoked in producing the final image which divides the cases. Where biochemical analysis is used the physical laws which describe the passing of information from strand to image are, naturally, biochemical and genetic laws, laws which are closely associated with, if not dependent upon, claims about DNA and the effects of closed loop DNA. But with the electron microscope the laws which trace the information are indifferent to DNA, its structure or the effects of being a closed loop. The laws of electromagnetism, electron scattering, etc., as listed above, are entirely independent of theories of DNA,[3] and for this reason the electron microscope offers the more reliable information of the shape of the DNA strand. It is not a matter of more or less theoretical support necessary to account for the information, nor the image being more or less noticeable to the untrained, which are relevant to the degree of observability in this case. Rather it is a question of more or less independence between the laws of observing and the theory of the object.

---

[3] In evaluating the independence of interpretation in this case one would want to distinguish those aspects T1 of the DNA theory which describe the global shape of the molecule and the effects of the shape, from those aspects T2 which describe the elemental constitution of the molecule. In this case it is T2 which is involved in accounting for informative interaction (scattering of electrons) and T1 which cares for the specific information that x is P (is a closed loop). The crucial point is that no DNA-theoretic claims about the molecular shape are involved to account for the scattering of electrons.

As a second specific example of electron-microscope observation, consider the account of 'Electron Microscopy of Individual Atoms' (Isaacson, *et al.* [1979], chapter 13 title). The microscope described by Isaacson is an STEM with a beam so narrowly focused that it scatters off one atom at a time. This eliminates effects of interference from multiple scattering centers. The imaging used is dark field imaging, hence there are no phase-coherence effects. The informational flow is quite simple in that an annular detector collects many scattered electrons when the beam is incident on a massive scattering center. The signal is sent to a television screen and in a position corresponding to the position of the incident beam a white dot is produced which is the image of the atom responsible for the scattering. In this way, crystal arrays, or even single atoms when they are alone in a crowd of much lighter atoms, can be shown. There are photographs of these kinds of results in Isaacson ([1979], p. 359). The picture of a 'single mercury atom' (figure 13.8a in Isacson) is completely described as a white spot on a cloudy dark background.

One can ask why this result is not described as the image of a nucleus rather than an atom. It is the nucleus, after all, which acts as the massive scattering center. The point is that the white spot more immediately represents the presence of an electron scattering center then it does an atom, and calling it information about an atom relies on a theory of atomic structure. The STEM probe and imaging are essentially just a compact system of Mott scattering (which is just Rutherford scattering with electrons rather than alpha particles as probes) where scattering information is correlated to the position in the specimen. This seems no more direct an observation of the atom than does some other phenomenon, such as beta decay, which is explained by atomic theory in terms of the presence of an atom. Both a beta-decay event and the location of a massive electron scattering center present the information 'that an atom is present' but in both cases the information is rendered only with the help of atomic theory. Imaging a single atom by its scattering information is like displaying a looped DNA through biochemical effects in that the lawful account that the final information is genuinely information about the object in question amounts to little more than an abductive inference of the presence of the object.

The important points that have emerged in this study of the observability of objects imaged by an electron microscope are these. The degree of observability of such objects depends on what it is and the specifics of its imaging. Certainly neither of these things, DNA nor single atoms, is observable to the unassisted human body. It is clear also that the imaging processes of an electron microscope are such that the observations are relatively more indirect than are observations with an optical microscope. In the optical case, the same photons which interact with the specimen can also interact with the human observer.

The analogous directness is not possible with the electron microscope as humans are not equipped to interpret information by electrons as they can information carried by groups of photons.

All cases of electron-microscope imaging involve massive support by auxiliary theories which describe the properties of electrons, their scattering, and their response to the electron-optical system. But it is not the amount of theoretical support which is distinctive in the different cases; rather it is the relation between that theoretical support and the theory of the specimen. It is a more reliable case of receiving information to rely on electron-atom scattering theory to probe the shape of a DNA molecule than it is to use the same to report on the presence of an atom.

5 THE DECLARATION OF INDEPENDENCE

The conclusions of the case of electron microscopic objects can be generalized to evaluate the epistemic significance of the dimensions of observability. What counts, according to the interaction-information account, is that information of the object-property pair and from the object to be conveyed to the scientist. In each case, information of the object is brought to the observer through causal interaction between the object and some medium of information, photons or electrons, for example, and subsequent causal interaction between medium and observer. Observation is an effect. The variation of directness represents the physical length of the causal chain, while the amount of interpretation indicates its epistemic length. But neither of these dimensions indicates the important distinction between the two examples of electron-microscopic objects. Those cases suggest that directness and amount of interpretations are relatively less important to evaluating observability than is the independence of interpretation.[4]

More needs to be said about independence of interpretation to make this important concept more precise and recognisable. Specifically, one wants a clarification of what it is for a theory T (of the object) to be independent of an observation account, that is, a set of theories (used to unnest the information). An initial evaluation of this kind of independence can be made simply by noting whether T (or that part T1 of T which mentions the object and property in question) is a member of the set of theories used in the account. If the object and its specific property must be mentioned to account for the flow of information (as in the biochemical evidence of closed-loop DNA) then there is a lack of independence. Further evidence of a lack of independence would be if any of the accounting theories relies on the observation for its confirmation.

---

[4] This, of course, is not to say that these other dimensions, directness and amount of interpretation, are to be ignored. A valuable interaction-information accounting must include an evaluation of these dimensions. But with the weighting of the independence of interpretation, and in the belief that this important dimension has been somewhat overlooked in past treatments of observability, I concentrate in what follows on that single dimension.

I should note, however, that Ian Hacking has made a similar point about the epistemic significance of independence of interpretation. Commenting on Shapere's example of using neutrinos to observe the interior of the sun, Hacking concludes, 'It is precisely the disunity of science that allows us to observe (deploying one massive batch of theoretical assumptions) another aspect of nature (about which we have an unconnected bunch of ideas)' ([1983], p. 183).

Still more work needs to be done on independence. Given that T is not an overt member of the set of accounting theories there is still want of a measure of independence (irrelevance) between T and each individual theory in the set. I do not have such a measure at this time. My hope is to motivate the intuition here, and soon to be able to make it more precise along these lines.

Since the information must be unpacked from effects, the reliability of the unpacking must be accountable, and that accountability comes as the independence between the causal laws used in analyzing the information, and the theory of the object itself. The epistemic reliability can come only with disinterest such that the theory describing observability has little or no important stake in the outcome. In this spirit, the laws of visual perception have no dependence on the kinematics of acceleration, and we are happy to acknowledge that the eye delivers acceleration information, hence the acceleration of midsized objects is what those with the urge to categorize would call observable. Similarly, electron optics and electron scattering theories do not rise or fall with the observability of DNA, and so they provide an independent accounting of the flow of DNA information to the observer. We are less willing to call the kind of effect produced in scattering electrons from single nuclei an observation of the atom, not because more physical interaction or more theoretical unpacking is required, but because the laws required to unnest the atomic information are atomic laws. It is the kind of nepotism in the accounting which impeaches the information.

6 INFERENCE TO THE SOURCE OF INFORMATION

The implication for scientific realism of the result of the independence-weighted interaction-information account of observability can be stated briefly as an inference to the source of information. This speaks to the specific issue of the realism of entities (as opposed to realism of theories), where the question is of the warrant for belief in the independent existence of entities as a function of their observability. When observability is described in terms of the information *from* x and of x, there is no sense to the discussion unless x exists. The reasoning here is closely analogous to Nancy Cartwright's 'inference to the most probable cause' ([1983], p. 92). The interaction-information account, like a causal account, has the existential claim built in in the sense that if you accept the fact that there is information about x which comes from x, then you also get the fact of the existence of the source of information.

Cartwright's discussion, and the inference to the source of information, insofar as the latter is modeled on the former, begin with the assumed acceptance of the account of the cause or source of information. Cartwright is very brief on the question of the epistemic markings of a good causal explanation. Her concern is only in pointing out the existential punch of 'the inferences you can make once you have accepted that [causal] explanation' ([1983], p. 93). She separates the issue of what should count as a good causal explanation from what you can do once you have one, by leaving the first problem in God's lap. If God tells us that (that is, we are certain that) quarks explain characteristics of hadrons, we are no closer to knowing whether or not quarks exist. But if God tells us that photons cause the retinal reaction of vision, then he has also told us that photons exist. And by interaction-information standards, if God tells us that DNA molecules are a source of information, then we learn as well that DNA molecules exist.

All this is good news as far as it goes, but of course God does not go around telling us these things. There is a serious incompleteness in the argument of realism with respect to entities x as causes of observable phenomena until we know *how to tell* that x is in fact the cause. The same obligation applies to the inference to the source of information. We now want to know what sanctions the inference about cause or the source of information so that we can measure the credibility of the corresponding existential claims and commit our belief accordingly. In the case of the causal explanation we must consult the relevant causal laws, and here it seems that the dichotomy between realism about entities (whether or not they exist) and realism about theories (whether or not they are true in what they say about unobservables) is blurred. To evaluate the inference to x as a cause, one must evaluate the truth of at least that part of x-theory which describes the causal powers of x. In other words, to complete the argument for the existence of the entity x (or for entity-realism in general) we must invoke an argument for the truth of that part of x-theory which portrays x in its role as a cause (in general, the causal accounting department of the theory must report truths). This much complicity between theories and entities must be admitted to complete the argument.

The interaction-information account of observability provides a mechanism to fill in the epistemic evaluation of the inference as to what is the source of information. As with a causal explanation, an account of interaction and transmission of information is done in terms of physical laws. The game is to consult the applicable scientific theories to describe observability. The best standard by which to measure such accounts is the dimension of independence of interpretation. An inference that x is a source of information is a good one insofar as the inference ticket is issued by an independent source, a theory with no stake in the specific properties or existence of x. Lacking the word of God, independence of interpretation is the next best measure.

Thus the interaction-information account with an emphasis on independence of interpretation provides criteria for evaluating claims about a source of information, and given a reliable specified source of information one can infer to the existence of the source of information.

7 SUMMARY

The interaction-information account is intended as a method for evaluating observability of scientific entities, using science itself as adjudicator of observability. The detailed features of the observability space are to be filled in under the influence of many case studies, and in this sense, the interaction-information account participates in an empirical approach to epistemology. The source of evidence for conclusions about observability in general consists of scientific theories and specific scientific accounts of observability.

The emphasis of the interaction-information account is to look for information of an object-property pair, information which is conveyed from the object to the scientist through interactions. The immediate payoff from the analysis is an appreciation of the complexity of observability space and the disparity of observability claims. Realization of the many degrees of freedom of observability is caution against reckless generalizing over unobservables. One finds enough difference between the observability of quarks and electrons, and even between that of quarks with respect to color and flavor, to realize that claims about microparticles as unobservables hide too much complexity to be useful in a debate about realism.

Beyond this caution against generalizations about unobservables, the interaction-information account has something to say about realism. This was reported under the heading of the inference of the source of information.

Clearly, the dimension of independence bears a heavy burden in the interaction-information account. The eclectic nature of observability evaluations, evidenced by the fact that there are few direct reports of observability in science because observability does not fall in the domain of any one theory, is insurance that the evaluation will not be an artifact of a theory. But could it be an artifact of a conspiracy among theories? Perhaps electron theory and electrodynamics take in the wash for biochemistry in the description of observability of DNA. Perhaps optics renders a similar service for kinematics in describing the observability of acceleration. Insofar as epistemology of science must be done from an internal perspective, that is, without prior information on the ontology of the physical world, we need to find criteria for epistemological evaluation within science itself. Lacking the word of God or advance knowledge of what there is in the world, independence in the interaction-information account is the surest epistemological foothold. In any case, the wash must be done, and in epistemology as in life, better some one else does it than do it yourself.

Department of Philosophy
Northwestern University

REFERENCES

CARTWRIGHT, N. [1983]: *How the Laws of Physics Lie*. Oxford: Clarendon Press.

DRETSKE, F. [1981]: *Knowledge and the Flow of Information*. Cambridge: MIT Press.

FODOR, J. [1984]: 'Observation Reconsidered', *Philosophy of Science*, 51, pp. 23–43.

HACKING, I. [1983]: *Representing and Intervening*. Cambridge: Cambridge University Press.

ISAACSON, M., OHTSUKI, M. and UTLAUT, M. [1979]: 'Electron Microscopy of Individual Atoms', in J. Hren, J. Goldstein and D. Joy (*eds*.), *Introduction to analytic Electron Microscopy*, pp. 343–68. New York: Plenum Press.

LEHNINGER, A. [1970]: *Biochemistry*. New York: Worth.

MAXWELL, G. [1962]: 'The Ontological Status of Theoretical Entities', in H. Feigl and G. Maxwell (*eds*.), *Minnesota Studies in the Philosophy of Science*, vol. 3, Minneapolis: University of Minnesota Press.

SHAPERE, D.[1982]: 'The Concept of Observation in Science and Philosophy', *Philosophy of Science*, 49, pp. 485–525.

VAN FRAASSEN, B. [1980]: *The Scientific Image*. Oxford: Oxford University Press.

Brit. J. Phil. Sci. 39 (1988), 469–494 *Printed in Great Britain*

Philosophy of Science: From Justification to Explanation

AHARON KANTOROVICH*

ABSTRACT

The paper investigates the implications of a nonaprioristic philosophy of science. It starts by developing a scheme of justification which draws its norms from the prevailing paradigm of rationality, which need not be universal or eternal. If the requirement for normativity is then abandoned we do not end up with a descriptive philosophy of science. The alternative to a prescriptive philosophy of science is a theoretical explanation of scientific decisions and acts. Explanation, rather than mere description, replaces justification; and the paradigm of rationality becomes a scientific paradigm. The implications of these results for the discovery-justification distinction are investigated. An explanatory philosophy of science deals with the generation, as well as with the selection of scientific conjectures; both contexts have an epistemic dimension.

Introduction
1 *Justification vs. Explanation*
1.1 *Justification and Paradigms of Rationality*
1.2 *From Description to Explanation*
2 *The Discovery-Justification Distinction*
2.1 *The Traditional D-J Distinction*
2.2 *Objections to the D-J Distinction and the Epistemic Dimension of Discovery*

INTRODUCTION

In recent years, the philosophy of science has undergone radical changes. With the decline of logical empiricism, it is not believed as widely as before that the source of scientific rationality can only be found in some system of formal logic or methodology. The philosophy of science, however, has not yet settled on a new, widely accepted path. Thus, fundamental questions are raised with respect to its scope, tasks, and methods. For example, what should supplement or replace the logical analysis of science? Should the philosophy of science be closely linked to the history of science or should it perhaps be converted into a science of science? There are however two prior, more fundamental, questions which have engaged traditional philosophers of science, and which are now posed more forcefully. The first question is whether the philosophy of science should adopt the task of *appraising* scientific claims; should it be content with the more modest aim of *describing* science, its methods and evolutionary patterns; or should it have both descriptive and prescriptive functions. The second fundamental question is whether the scope of the pilosophy of science should be restricted to the so-called 'context of justification' or should it also deal with the 'context of discovery'. The distinction between the two contexts is also questioned. The prescriptive-descriptive dichotomy and the discovery-justification distinction evoke the cardinal issues related to the nature of the philosophy of science.

---

* This paper was written under the auspices of the Wolfson Chair Extraordinary of Theoretical Physics, Tel Aviv University.

We will start by analysing the prescriptive-descriptive dichotomy and arrive at the conclusion that neither purely prescriptive nor purely descriptive philosophy of science are possible. We will devise a nonaprioristic scheme which can be viewed as a scheme of justification or as a scheme of explanation. As a scheme of explanation it will function as a science of science. Explanatory philosophy of science does not undertake the task of telling scientists how to do science. We will argue that it can serve instead as a 'therapeutic' aid for scientists.

These results will bear upon the discovery-justification distinction. Traditional philosophy of science deals only with the context of justification or evaluation, leaving the context of discovery to a descriptive science such as psychology. In our scheme of justification the context of evaluation, too, is amenable to scientific investigation. In fact, the entire distinction becomes blurred.

I JUSTIFICATION VS. EXPLANATION

1.1 *Justification and Paradigms of Rationality*

1.1.1 Science as a Goal-Directed Activity

The main stream of traditional philosophy of science starts off with a normative or prescriptive attitude. It seeks rationality in science, *i.e.* it looks for the logic and reasoning behind scientific acts. Scientific rationality depends on the goals of science. It is therefore the first task of the philosopher of science to uncover these goals. If the goals are known, the philosopher can try to answer the question as to whether or not the proposed means for achieving them are appropriate—*i.e.* rational. Most traditional philosophers of science have taken for granted the assumption that the main goal of science is reaching comprehensive truth about the world. The goal of explaining natural phenomena is related to the above goal.

Truth is a property of statements. Therefore, assigning to science the goal of truth means that the task of science is to generate true statements about the world. Hence, the rules of propositional calculus or predicate calculus, of classical logic, are the natural candidates for showing us how to do good science. If we see the task of science as generating statements which are highly probable rather than true, then some theory of probabilistic inference will guide us in doing science rationally. Thus, deductive, inductive or probabilistic inference schemes will be the basis for rational acts in science. Adopting this approach, which may be called 'logicism', the philosopher of science views science as an 'inference machine'.

Logicism has faced insurmountable difficulties. Some of these difficulties are related to the impossibility of reconciling many of the actual acts and decisions of scientists throughout the history of science with those recommended by logicist methodologies. One could assume that this might lead philosophers of science to review their fundamental presuppositions. One such presupposition is that science is a truth-seeking system.

If we abandon this presupposition while still believing that science is a goal-directed activity, we can try to suggest alternative goals. We may do this by examining the declarations of the scientists themselves throughout the history of science. We will find out, indeed, that there are other declared goals besides the goal of truth. For example, the following three goals are very often cited:

(a) The goal of predicting natural events and phenomena.
(b) The goal of constructing an economical theoretical system which will entail our observational statements.
(c) The goal of advancing technology and mastering nature.

A more radical approach is to look for the goals of science in the realm of the subconscious, *i.e.* to look for collective motives which scientist are not aware of. Scientists declare that they seek comprehensive truth, objectivity, etc., but their real motives may be psychological or social. For example, John Ziman suggests [1968] that the goal of the scientific community is to arrive at a consensus, rather than truth; this goal of consensus is what distinguishes science from other human activities. This approach treats science as a social or even biological phenomenon. Scientists are not free to choose their goals; they can only choose to participate in the process and obey the rules of the game, without being fully aware of its significance. This being the case, the activity of the individual scientists cannot be judged by a standard of rationality, as if it were a goal-directed activity. The individual soldier is not always aware of the goals of the army as a whole; for example, he may get an order to retreat while the army advances.

1.1.2 The Dilemma of the Normative Methodologist and Goodman's Solution

The paradigmatic model which guides the normative philosopher of science with the goal of truth in mind, is deductive logic. Logic yields prescriptive criteria for deductive validity of reasoning. These criteria do not attempt to describe how people *actually* reason but how they *should* reason. When people violate the rules of logic the logician would say that they are in error. Similarly, criteria of scientific rationality are not intended to be descriptions of how scientists actually reason, but rather how they should reason.

There arises the question from where does the normative philosopher of science take his rules of rational reasoning, and how does he justify them. If his theory of rationality is derived from a first philosophy or *a priori* principles, such as the principles of logical reasoning, then the problem of possible violations of his recommendations in actual science will arise. He would face serious difficulty if, in the light of his first philosophy, he recommended abandoning some of the central ingredients of scientific practice. Such science could turn into philosophically fabricated science; the most celebrated successes of actual science might not have been achieved had scientists adopted his methodology. Thus, the philosopher who attempts to deal with real science and not with an ideal system of reasoning must keep an eye on the history of science. On the other hand, as a normative philosopher he should justify the methodological rules he prescribes, in light of the goals of science. Thus, the normative philosopher of science faces a dilemma: on the one hand, he wishes to maintain the notion of justification; and on the other hand, to avoid a situation where justified rules of inference are systematically violated by most scientists most of the time. He has to choose between abandoning his first philosophy or rejecting most of the celebrated chapters of science as irrational or nonscientific. For example, an empiricist philosopher may adopt the epistemological view that only observation sentences are justifiable. He may draw from this the methodological rule that only theories which are wholly reducible to observation sentences are scientific. If he then finds out that Newtonian mechanics or quantum mechanics cannot be wholly reduced to observation sentences, he must either conclude that modern physics, which is erected upon these theories, is nonscientific or nonrational, or abandon his first philosophy as a theory of rationality. The first possibility is bad since it means killing science altogether. The second is worse since a first philosophy is, by definition, irrefutable by facts.

Nelson Goodman provides us with an escape from this dilemma. With his approach we also avoid the task of finding out what the goals of science are; the justification of methodological rules is not dependent on the goals of science. Goodman starts with an analysis of justification of deductive rules. This analysis prevents the possibility of a fundamental disagreement between logical theory and common sense arising:

> Principles of deductive inference are justified by their conformity with accepted deductive practice. Their validity depends on accordance with the particular

> deductive inferences we actually make and sanction. If a rule yields inacceptable inferences, we drop it as invalid. Justification of general rules thus derives from judgements rejecting or accepting particular inferences ... A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend. ([1965], pp. 63–4)

In other words, the justification of rules of inference is not based on *a priori* principles, but on their accord with inferential practice. This scheme contains a prescriptive element since 'an inference is rejected if it violates a rule we are unwilling to amend.' Note that the reason for this unwillingness is not specified; it is treated as a given fact. Stephen Stich and Richard Nisbett describe this sort of accord by saying that a justified rule is in 'reflective equilibrium' with inferential practice ([1980], p. 190); they borrowed the term from John Rawls ([1971], p. 20). We would add that this equilibrium should be dynamic, if we wish to entertain the possibility of changing our attitude towards rules and particular inferences as our reasoning experience evolves.

Goodman generalizes this view of justification to include inductive reasoning: rules of induction are justified by their being in reflective equilibrium with inductive practice. We may further generalize this analysis to include all methodological rules in science; for example rules of confirmation or falsification, or acceptance or rejection, of scientific theories and observational statements. Methodological rules may be rules of inference or rules which guide decisions, *e.g.* decisions to accept or reject theories, decision to perform certain observations or experiments, etc.

This approach to methodological rules does not presuppose any particular goals for science, or that science is a goal-directed activity at all. Its validity merely derives from its accord with scientific practice. If, however, we assume that science is goal-directed, then the line of reasoning with respect to the goals of science is reversed here: when a stable set of rules is found to be in reflective equilibrium with the inferences and decisions of a given community (*e.g.* the whole scientific community or a community of scientists engaged in a specific branch of science) we assume that rational behavior in that community means obeying these rules. This means that rationality is not universal but is community-dependent. Given the rules, we can infer or hypothesize what possible goals the community attributes to science. For example, if we find that physicists are guided by a methodological rule which requires conducting active experimentation rather than making only passive observations we may come to a conclusion that one of their goals is to reproduce and control natural phenomena, or to advance technology (Kantorovich, [1982]). The goal of attaining true descriptions and explanations of natural phenomena cannot by itself explain why, for example, particle physicists produce more and more new, short-lived particles at higher and higher energies. It seems that by using this method of active research, physicists create artificial phenomena rather than discover natural phenomena. We can imagine a science which would not intervene in the natural course of events and would only be engaged with recording natural phenomena. Indeed, such a passive approach was actually adopted before the emergence of modern experimental science. Thus, the goal of controlling natural phenomena and advancing technology can be seen as one of the goals which distinguishes modern natural science from its predecessors.

1.1.3 Justification and Explication

Goodman's approach to justification is closely related to the notion of explication of intuitive rules. Carnap introduced this notion while attempting to explicate various concepts of probability and induction. By 'explication' he meant formalization or axiomatization. Explication transforms vague concepts used in ordinary or scientific discourse into clearer concepts the rules of use for which become rigorous. A good explication should achieve a good adjustment between the formal system and the intuitive concepts. Such a process can expose and remove inconsistencies in the use of the intuitive concepts.

Mary Hesse advocated reducing the problem of justification of induction to the problem of explicating intuitive inductive rules. According to Hesse the process of explication is divided into two tasks:

> (i) To formulate a set of rules which capture as far as possible the implicit rules which govern our inductive behavior. (ii) To formalize these in an economical postulate system ([1974], p. 97)

An example for a rule of induction generated at stage (i) is the rule of enumerative induction, which can be formulated as follows: 'An empirical generalization should be increasingly confirmed by observation of an increasing number of its positive instances and no negative instance.' Another example for a presystematized rule is the following rule which Hesse subsumes under the category of induction, but which can just as well be classified into the hypothetico-deductive method: 'A hypothesis should be strongly confirmed by the observation of the truth of one of its logical consequences which was not expected before the hypothesis was proposed.' A formal system most appropriate for explicating such rules of scientific method is probability theory. In such a system the probability of a theory, for example, explicates its degree of confirmation.

Let us now compare Hesse's two-stage scheme with Goodman's notion of justification. At first sight we might be tempted to identify the process of reaching reflective equilibrium with stage (i); that is, the presystematized rules of inductive inference are perhaps formulated via an interaction with inductive practice, and thus justified. However, Hesse seeks justification specifically for the postulates of the *formal* system generated at stage (ii) and not just for the inductive rules formulated at stage (i):

> a sufficient, and perhaps the only possible, justification of a set of postulates of inductive inference would be that they form a 'good explication' of the intuitive inductive rules. Justification in this sense resides in the interaction of postulates and rules and not in any external support for the postulates independently of the rules. (ibid, p. 98)

Goodman refers to two methodological levels: (a) particular inferences, or inferential practice, and (b) principles or rules of inference, where the latter need justification. Hesse refers to three levels: (a′) implicit inductive rules; (b′) an explicit set of rules; and (c′) a formalized system, where the postulates of the latter need justification. The following table lists the methodological levels of the two schemes.

Goodman	Hesse
(a) particular inferences	(a′) implicit inductive rules
(b) rules of inference	(b′) explicit inductive rules
	(c′) formal system

The implicit rules of level (a′) govern the inferential practice of levels (a). Furthermore, levels (b) and (b′) are identical. We can therefore identify stage (i) of the explication process with the process of reaching reflective equilibrium between rules [level (b)] and practice [level (a)]. This process, according to Goodman, provides justification to the rules. According to Hesse, however, the justification is shifted 'upwards' to the formal system [level (c′)] and it is attained through the interaction of the system with the rules.

1.1.4 Paradigms of Rationality

I now propose a scheme of justification which employs the major elements from the above two approaches to justification. The central role in this scheme will be played by a notion which I call '*paradigm of rationality*.' This notion will contribute to the normative dimension of the scheme. As we will see, it is akin to one of the uses Thomas Kuhn makes of his notion of paradigm. The paradigm of rationality replaces the first principles of rationality of the aprioristic philosophy of science. However, unlike the latter the paradigm of rationality is not eternal, and may undergo changes with the evolution of science.

The structure of our scheme of justification can be represented by analogy with the structure of theoretical explanation peculiar to modern science. We can describe such a process of explanation as an interplay between the following layers of scientific knowledge: (a) observational data, (b) empirical generalizations, (c) an explanatory theory and (d) the general world picture prevailing in science. For example, the kinetic theory of gases [layer (c)] explains the empirical laws describing gas behavior, such as Boyle's law, Gay-Lussac's law and the ideal gas law [layer (b)], which 'summarize' the observational data [layer (a)]. According to the view of explanation which I will adopt here, not every theory which entails empirical generalizations explains them; *a necessary condition for a theory to be explanatory is that the theory comply with the world picture.* For example, an explanatory theory in 19th-century physics had to comply with the mechanistic-corpuscularian world picture [layer (d)]. The kinetic theory not only entailed the gas laws, but also explained them, since it complied with that world picture. Furthermore, *only empirical generalizations which are faithful to the observational data are candidates for explanation.* In the context of scientific explanation this requirement seems trivial; no one would suggest explaining 'laws' or 'generalizations' which contradict most of the data. An explanatory theory, however, may somewhat correct the original generalizations. For example, Newtonian theory entailed a corrected version of Kepler's original laws of planetary motion which it intended to explain. Furthermore, the explanatory theory may, or should, predict new laws. If the modified generalizations or the new laws agree with the data, the theory is strongly confirmed.

Our scheme of justification can likewise be divided into four layers respectively: (a) particular scientific inferences and decisions ('scientific practice'); (b) methodological rules; (c) a methodological theory, which may be a formal system (such as Hesse's) but not necessarily; (d) the paradigm of rationality. According to the conception of justification I propose here, *a methodological theory justifies the methodological rules which it entails only if it complies with the paradigm of rationality.* This is our first condition for justification.

Before trying to characterize in general the notion of 'paradigm of rationality' let us illustrate its meaning by means of some examples. First, let us consider *logicism.* If we do not treat logicism as *a priori* valid, we may treat it as one possible paradigm of rationality. This paradigm views science as proceeding by inferences. The goal of science is generating true, or approximately true, statements or to eliminate false statements. In order to achieve this goal, scientists should obey prescribed rules of inference. This view of science would require the construction of a formal system analogous to, or an extension of, deductive logic. Possible methodological theories which comply with this paradigm are inductive logics, probabilistic confirmation theories such as Bayesian theory or falsificationist methodology. (Falsificationism may alternatively be subsumed under the paradigm of 'evolutionism' which will be mentioned below.) Hesse's scheme may be viewed as such a methodological theory, provided the requirement for a formal system is not treated as an *a priori* requirement.

Another example is *sociologism*, which views science as a social phenomenon. This view may have a number of versions: *e.g.* science as a tool for the advancement of society, or in particular of technology, or science as a tool for achieving particular social goals. If the goal is advancing technology, a methodological theory should imply rules of preference for research topics or rules for choosing between theories according to their technological utility.

Finally, let us consider the paradigm of *evolutionism*, *i.e.* science as an evolutionary phenomenon analogous to, or constituting a continuation of, organic evolution. One possible methodological rule which might be mentioned in this context is the need for proliferation of hypotheses such that there will be great variability, this being the source of evolutionary progress. Another possible methodological recommendation derived from this paradigm pertains to the manner by which hypotheses should be generated, *i.e.* independently of the data and of the problems to be solved (blind variation). Finally, rules of elimination and falsification (selection) are indispensable to this paradigm.

How we can generalize from the above examples in order to characterize a paradigm of rationality? First, it should be noted that in each of the above cases there is more than one methodological theory which complies with, or is implied by, the paradigm of rationality. Hence, a paradigm of rationality cannot be equated with a methodological theory. Second, each of the above paradigms says something about the nature of science and its goals; each sets a general theme for science: 'science as an inference machine', 'science as a social phenomenon' or 'science as an evolutionary system'. Third, since the paradigm of rationality specifies the goals of science or its general nature, it implies general requirements to be met by a methodological theory.

Our second condition for justification is that *the justified methodological rules be faithful to scientific practice*, *i.e.* that they be in reflective equilibrium with scientific practice. In other words, our notion of justification can be applied only to rules which accord with scientific practice. This is analogous to the notion of explanation which can be applied only to empirical laws or generalizations which accord with the observational data. This condition guarantees that we will not generate justified methodological rules which are not adhered to by practicing scientists. Philosophers of science who base their methodology on a first philosophy *do* generate justified rules which are not obeyed in actual science, hence the above condition is not trivial as is its counterpart in the scheme of scientific explanation.

Our scheme, however, is normative or prescriptive. As in Goodman's approach, we can sometimes refuse to amend a rule which does not accord with a particular scientific inference or decision. Our paradigm-guided scheme gives us a definite reason for refusing to amend a rule when it is entailed by an established methodological theory which complies with the paradigm of rationality. Therefore, when a particular inference or decision violates such a rule, we will reject such an inference or decision. However, a paradigm of rationality which leads us to rule out some important particular inferences and decisions, or a large proportion of scientific practice, should in the long run be amended or be replaced by a new one.

Thus, the normativity of our scheme is expressed by the fact that it may lead to the rejection of particular inferences or decisions, and by the fact that it lends justification to particular inferences or decisions which comply with the methodological rules. However, this is not an aprioristic justification, since the methodological rules have themselves been justified by being in reflective equilibrium with scientific practice and by complying with the paradigm of rationality.

According to this approach the system of methodological rules is dynamic; it may incorporate new rules or drop old ones, according to the changing practice of science, on the one hand, and the changing image of science and scientific rationality, on the other hand.

Hence, the standards of rationality are not dictated entirely from outside science by some first philosophy. They are rather drawn from both the practice of science and the paradigm of rationality. The paradigm of rationality in its turn interacts with scientific practice through the methodological rules. Let us give a hypothetical example for the interaction between the different layers of the scheme. Let us assume that we start with a logicist paradigm of rationality. At a later stage considerations of technological utility may infiltrate the paradigm of rationality and give more weight to theories which lead to technological advance. At this stage new modes of inference or methodological rules may appear which lead to successful technological inventions. For example, if according to the logicist approach two competing theories have equal methodological merits, but only one of them leads to a discovery which opens the way to a useful technology, the latter may be considered to be confirmed to a higher degree. If it happens that these modes of inference do not comply with the existing logicist rules, they may nevertheless be assimilated into the system of methodological rules, reinforcing the technologically oriented trend of the paradigm of rationality.

1.1.5 A Historicist Paradigm of Rationality

We may come to the conclusion that certain celebrated chapters in the history of science set up novel standards of rationality which cannot be reduced to one of the above-mentioned paradigms of rationality and their like. We may base our decision as to what chapters to choose as paradigmatic cases on the judgements of scientists; we can look for the paradigmatic cases on which scientists model the methods and procedures they adopt in conducting their research.

One might adopt, for example, the Newtonian model of scientific research as a paradigm of rationality. The methodological theory which might be drawn from this model may require the search for dynamical laws which explain empirical laws. Here again, there may be different methodological theories which can be drawn from this paradigm. Thus, instead of classical logic as a model for science we have a specific chapter from science as a model. If this model yields results which do not accord with some scientific practice which we do not want to reject, the paradigm of rationality should be modified. Perhaps we will be led to the conclusion that the standards of scientific rationality change with the history of science, possibly with some hard core remaining unchanged.

Indeed, more and more philosophers of science who have freed themselves from the logicist paradigm are ready to learn from the history of science. They try to construct theories of science which draw lessons from the history of science. Imre Lakatos' methodology of scientific research programs (Lakatos, [1970]) and Larry Laudan's problem-solving approach to scientific rationality (Laudan, [1977]) are examples for a methodological theory which is strongly inspired by case studies from the history of science. Thus, the philosopher of science who adheres to this approach would expect that scientists be guided by a paradigm of rationality modelled on past achievements of science. The nonreflective scientist may not be aware of the fact that he is imitating certain models of doing science, or of what models he is imitating. The philosopher of science, from his external vantage point, may expose the model imitated. Having a historical perspective, he will discover that this methodological pattern has characterized science throughout its history.

The following are examples of methodological models which have been repeatedly followed in physics: (a) the use of explanatory theories; (b) the procedure of starting with an ideal or simplified model and adjusting it to the data (examples are the kinetic theory of gases, the Bohr–Rutherford atomic model or the early quark model); and (c) the general procedure of drawing analogies with known phenomena or entrenched theories. The viability of this last methodological practice can be illustrated by a recent example, namely the analogy made between the nuclear force and the interatomic force:

> The electromagnetic force binds electrons and nuclei to make atoms. The atoms, although they are electrically neutral, interact through a residual electromagnetic force to form molecules. The strong force binds quarks to make protons, neutrons and all other hadrons, and the residual strong force between protons and neutrons is the so-called nuclear force that binds them into nuclei. (Haber and Kane, [1986]).

According to the above view scientists are not engaged merely in learning facts about the world, but also in learning how to learn, or learning how *science* learns from experience. Science produces, therefore, a second-order knowledge, *i.e.* knowledge about scientific knowledge. This is exactly what is done in machine learning. Expert systems, for example, are fed by information drawn from experts. This information summarizes not only their factual knowledge, but, more importantly, the rules which guide their problem-solving activity. These rules, which are nonformal and intuitive, constitute the *heuristic* of the given field of knowledge. An historicist or internal paradigm of rationality, likewise, serves as a kind of a heuristic which is drawn from the practice of leading scientists in successful phases of the history of science. There may be different heuristics, *i.e.* paradigms of rationality, in different branches of science, and in different periods.

Let us call this general conception of rationality 'the historicist conception of rationality'. According to this conception, the paradigm of rationality which guides scientists at a given period is a product of a long-term historical process. The historicist conception of rationality is still normative, even though to a minimal degree; the sources of normativity are selected chapters from the history of science.

In sum, the nonaprioristic approaches to scientific methodology take into account scientific practice which should be in reflective equilibrium with the rules derived from the methodology. In the historicist case the history of science is reflected also in the paradigm of rationality. Thus, the dichotomy between prescriptive and descriptive methodologies becomes blurred; a purely prescriptive philosophy of real science is impossible. In the next section we will see that a purely descriptive philosophy of science, too, is impossible.

1.2 *From Description to Explanation*

1.2.1 The Point of Departure of a Descriptive Philosophy of Science

In the previous section we described a philosophy of science the starting point of which is normative or prescriptive. We will now consider the implications when the point of departure is descriptive. First we have to see how a descriptive philosopher of science is different from a historian of science. A historian of science cannot be a neutral observer and describe the 'mere' facts since he has prior expectations and attitudes towards science, and he has initial concepts by which he comprehends the phenomena of science. As Lakatos puts it: 'history of science without philosophy of science is blind'. (Lakatos, [1971], p. 91) What then is the difference between the two, if any? A minimalist historian of science might be distinguished by his intention to be as 'neutral' as possible, *i.e.* to describe the empirical facts and to avoid using laws or theories as far as possible. On the other hand, the descriptive philosopher of science, being a philosopher or a methodologist, expects to find in science methods and general characteristics. He expects science to be a rule-governed phenomenon. Such a position involves an intention to analyse, to generalize or to theorize and not just to remain on the level of reporting what scientists do. However, since he does not intend to be prescriptive, he would not adopt a scheme of justification such as those described above.

1.2.2 The Problems of the Descriptive Approach

The descriptive philosopher of science will not find any list of explicit and clear methodological rules which guide scientists in their work. Although some general methods of science are occasionally discussed in scientific literature, there is no general agreement with respect to their clear formulation. Furthermore, scientists do not learn their profession by studying a methodology. One of the lessons which a graduate student learns when he turns to actual research is that he has to ignore many of the nice and neat principles and slogans he has learnt during his undergraduate studies, in particular some of those principles which are supposed to constitute the scientific method.

Even when great scientists mention certain methodological principles, the philosopher of science may find that the scientists do not actually adhere to them. Perhaps the most conspicuous example of this appears at the very beginning of modern science. Isaac Newton, who declared 'hypotheses non fingo' or 'I feign no hypotheses', created one of the most celebrated hypotheses in the history of science. Newton's theory of gravitation and motion goes far beyond common-sense experience, and has far-reaching predictions, some of which were not dreamt of by Newton's generation. Newton's use of the term 'hypothesis', however, is different from that of 20th-century physicists and philosophers of science. In one case, for example, he uses this term to mean a proposition which refers to 'occult qualities' which are not observable or measurable. This indicates another problem with which the descriptive philosopher of science is faced. Contemporary examples are abundant. A typical example is that of the theoretical physicist who emphatically declares that his theories are nothing but an economical means of organizing observational data. The philosopher of science might point out that such a physicist employs the hypothetico-deductive method, where the theory goes beyond a mere summary of observed data. Another example is that of the scientist who claims that he is making observations in order to confirm a theory, but a philosopher (such as Karl Popper) might tell him that his experiments are actually attempts to refute the theory. Another philosopher (such as Thomas Kuhn) might tell him that he is just solving problems in the framework of a normal science.

Thus, the philosopher of science who views his task as descriptive, faces the problem that he cannot take at face value the declarations of scientists about the scientific method in general, and even about the principles which they employ in their own research. So perhaps a descriptive philosophy of science should not take very seriously what scientists say, but rather study how scientists actually *do* science. For example, the descriptive philosopher of science should study how scientists construct theories and check them against experimental results, and then he might try to analyse and generalize from his findings. However, here arises a problem which faces the historian of science:

the philosopher of science does not encounter neutral facts when he studies science; already when he starts his study he has to choose where to look and how to interpret and categorize what he sees. By the metascientific terms he employs he tries to capture the data of science. The raw data may include scientific papers and reports, or conference proceedings, or letters (scientific products) or perhaps more abstract entities such as theories and experiments. Here there is a parallelism between a descriptive philosophy of science and science itself. It is a widely accepted view that there are no pure observational terms and statements in science; every descriptive statement employs theoretical terms and is loaded with theoretical assumptions. The same applies to descriptive statements *about* science.

The descriptive philosopher of science will naturally try to first use the metascientific terminology employed by the scientists themselves. As in the case of the scientists' declarations about the scientific methodology, however, he will very soon find out that there exists no such unified and consistent terminology. As we have seen from the example of Newton's use of the term 'hypothesis', metascientific terms may be interpreted in different ways at different times. Even in contemporary scientific writings, we do not find any systematic metascientific terminology. Metascientific terms as used by scientists are frequently ambiguous. The term 'science' itself conveys different meanings to different scientists. Terms such as 'theory' or 'model' are used with a variety of meanings. The term 'theory', for example, which is central to modern science and which is extensively used by scientists and philosophers, has a number of possible meanings, some of which are interrelated. A theory is: (1) a conjecture, as opposed to a solid factual statement; (2) a system of statements which employs so-called 'theoretical terms', *i.e.* terms which do not appear in the observational vocabulary; (3) an explanatory system, as opposed to an empirical generalization which does not explain but only describes and summarizes observational data; (4) a system of laws of nature; (5) an uninterpreted deductive system which is related to observational data through correspondence rules; or (6) a dynamic system which is not a system of statements at all, but which generates a succession of theory versions (statements) throughout its history. The historical entity called 'Newtonian theory', stretching from the 17th century till the end of the 19th century, exemplifies this last interpretation. Furthermore, terms such as 'theory' and 'model' are sometimes used interchangeably to refer to the same entity, *e.g.* the Bohr atomic model or theory. Methodological terms such as 'proof' and 'refutation' are frequently used misleadingly: scientists often claim that a certain theory was proved or refuted by experiment, whereas it is well-known that if theories are universal statements, they cannot be logically proved by any finite quantity of observational data. It is further known that logical refutation can be avoided by making *ad-hoc* modifications of the theory or the data, or by introducing some auxiliary assumptions.

Thus, the descriptive philosopher of science must choose for himself a proper metascientific terminology and a proper categorization of scientific activity and products. He may use in a more refined manner some of the terms employed by scientists. He may reinterpret other terms and add new ones; the terms 'research program' and 'paradigm' are examples of the latter. The choice of terminology and categorization will be made in compliance with criteria which guide the philosopher of science, such as fruitfulness or explanatory power.

1.2.3 Descriptive Philosophy of Science as a Science of Science

In other words, the descriptive philosopher of science is led to play the role of a scientist; he will invent a scientific theory of scientific method or a scientific theory of science. If the term 'theory', for example, appears in such a metascientific theory, it will be a theoretical term.

When descriptive philosophy of science becomes a 'science of science' it may adopt some of the methodological principles it finds in science. The circularity of this situation leads to some questions. When the scientist of science starts investigating the methodology of science he has to start with some initial methodological rules which will guide his own investigations. What rules, however, will guide him before he has found what the rules are? The answer is that he will conduct his investigations as every scientist does—without necessarily knowing explicitly what the rules are which guide him. If there are any explicit methodological rules to which he adheres, they will be included in his initial hypothesis about the methodology of science. In any case, just as any scientist, he cannot start his enquiry without being equipped with prior views and assumptions, or with a preliminary, semideveloped theory. He will start his own investigations by using some of these methods. Later on he may change his views as a result of his investigations, with possible implications for his own methods.

It is not necessary, however, that the same methodological rules which the scientist of science claims to be effective in science will be the ones which guide his own investigations. For example, he may reach a conclusion that the methodology of science is pluralistic, and that the human or social sciences employ different kinds of methods than do the exact or natural sciences. If the science of science belongs to the group of human sciences and if the main concern of the scientist of science is to study the methodology of the exact natural sciences, then there will be no circularity here.

In the event that he concludes that there is a methodological core common to all sciences, the scientist of science should start with some methodological rules which, according to his hypothesis, guide all scientists. For example, if he offers a theory of confirmation which is supposed to be common to all sciences, this theory should be confirmed by its own standards. There is no self-supporting circularity here, for it is entirely possible that a theory of confirmation may not be confirmed by its own standards. For example, suppose that according to a confirmation theory Tc, a necessary condition for a significant confirmation of a scientific theory is the occurrence of a successful prediction of an event or a phenomenon which had been unknown and unexpected before the theory was invented. The theory Tc itself will not be confirmed before the philosopher of science finds a previously unknown event in the history of science which is predicted by Tc. It would lend confirmation to Tc if there were a scientific theory which explained beautifully a vast amount of known data and phenomena, with none of its predictions having been found to be false, but which has not been accepted because it has not predicted any novel phenomena.

The circularity which may arise in a scientific investigation of scientific methods would be vicious were we to claim that our methodological theory gave absolute justification to some methodological rules. Since our approach includes no such claim, the danger of vicious circularity is avoided.

1.2.4 Explanation Instead of Justification

If we treat our theory of science as an explanatory theory, it should explain the behavior of scientists, their inferences, decisions and acts. It may explain, for example, why scientists accept or reject a theory in a given situation—a situation which might be characterized epistemologically, sociologically or psychologically. (Epistemology is regarded here as naturalized, *i.e.* as part of psychology—as Quine [1969] would have it—or even part of biology—as evolutionary epistemologists maintain. Hence it can be treated on a par with sociology and psychology.) It may explain *why* scientists perform certain experiments and *why* they are investigating a certain subject. This is in contradistinction to a normative theory of science which would prescribe which experiments *should* be done, which subjects *should* be investigated in certain methodological situations or which decisions are justified.

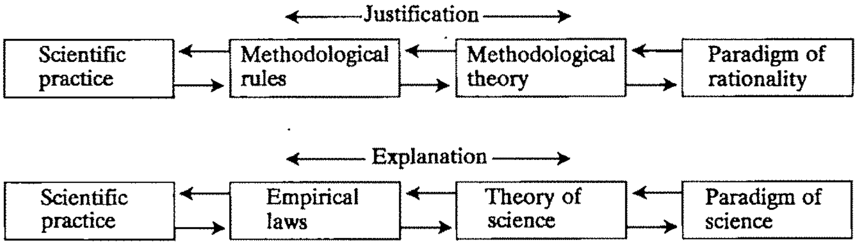
When the normative philosopher of science observes a case where his prescribed rules are violated by scientists, he says that the scientists are acting irrationally, or that they are committing a fallacy. In Goodman's approach such a fallacy is relative to the collective practice of the scientific community as it is reflected in the rules. In our scheme the fallacy would be regarded also as relative to the paradigm of rationality. A parallel situation would arise in the *scientific* approach to methodology when scientists violate, by their decisions or actions, certain implications of an entrenched theory of science. The scientist of science would not say that these scientists are in error, but rather that the scientists in question are in deviation from the accepted patterns of behavior predicted by the theory. The theory may then predict that the scientific community will not accept their results or findings, or even totally ignore those scientists, If this prediction is violated, or if the majority of the scientists in a given field at a given period of time violate the rules or the patterns of behavior predicted by the theory of science, then such a situation would be treated as an anomalous phenomenon. The anomalous behavior should be explained by the theory in conjunction with some relevant information about the conditions under which these particular scientists (which may constitute the whole scientific community at a given period) are operating. This information may refer, for example, to psychological factors or to external social or political forces bearing upon the scientists. The explanation may require the introduction of some auxiliary assumptions; it all depends on the sort of methodology of explanation adopted by the theorist of science. The situation which faces the philosopher of science in this case resembles that of an anthropologist investigating a remote society or an unfamiliar culture. In his investigations he may develop a theory about the phenomena he encounters. The theory may be derived from his general view about the nature of societies and cultures, and from his observations in the field. The theory may then yield empirical predictions. If the anthropologist encounters violations of his empirical predictions he does not reject the behavior of the members of the society he investigates, but rather attempts to explain the violation.

There is one sense, however, in which the science of science may fulfill a prescriptive, though not a normative role. When the theorist of science has a successful explanation for 'deviant' behavior on the part of scientists this means that he knows which conditions caused the anomalous behavior. He may advise them to change their behavior by avoiding the problematic conditions and reintegrating with the 'normal' course of science. Scientists may well desire to receive such advice, for their 'anomalous' behavior will cause them problems, such as failure to achieve consensus in the community, or to arrive at explanations which are satisfactory according to their own standards. The status of such an advisory role is analogous to the status of psychotherapy, which attempts to overcome or avoid the conditions which are disturbing the patient. The psychologist does not claim any status of normativity; he does not consider the abnormal or disturbed pattern of behavior as *bad*. Psychological theory would only predict that in the abnormal course of behavior it is more likely that a person or a community encounter more problems, psychological and social. Thus the theory is prescriptive relative to a putative desire of a person or a community to avoid such problems.

The prescriptive element in the science of science is therefore not accompanied by a claim for normativity or justification, just as 'normal' emotional behavior is not claimed to be more justified than 'abnormal' behavior. Actually the advice which can be given to scientists is given by an *applied* branch of the science of science which serves scientists in response to their own desires, just as clinical psychology is an applied branch of psychology.

We can still employ our four-layered scheme to describe the structure of a descriptive philosophy of science as a science of science. The scheme of justification will turn into the original four-layered scheme of explanation which we have borrowed from science. Thus, scientific practice will be treated as the layer of empirical data. The methodological rules will become empirical laws. The methodological theory, or the theory of science, will become the explanatory theory. Finally, the paradigm of rationality will turn into the philosopher of science's general outlook on science, *i.e.* his Kuhnian paradigm, through which he views science as a phenomenon. The empirical laws, which include the methodological rules, summarize and generalize scientific practice. The theory of science, which complies with the paradigm, explains both rules and practice.

The same four-layered structure can be viewed as a scheme of justification or a scheme of explanation, depending upon the meaning attached to the paradigm. If the paradigm is a paradigm of rationality, we are in the realm of justification; if the paradigm is our general view of science as a phenomenon, we are in the realm of explanation. The dual structure is exhibited in the following diagram:



Let us illustrate this duality between justification and explanation with respect to logicism. When we treat logicism as a paradigm of rationality we mean that particular scientific inferences and decisions are justified only if they obey the rules prescribed by the methodological theory. If we change our attitude towards logicism and treat it as a scientific paradigm, this means that it is used as a guide for constructing explanatory theories for the science phenomenon. In this capacity, logicism may be a general psychological paradigm on the nature of scientific knowledge, and a *general* view of how scientists reason in fact, rather than a normative view of how they *should* reason or act. Thus, naturalized epistemology becomes part of psychology. The task of the epistemologist qua scientist is to propose hypotheses as to exactly what the rules of inference are. A theory which explains and describes the reasoning and action of scientists is an empirical theory which is testable and refutable. Violations of the empirical rules of inference will be treated as problems to be solved or anomalies to be explained.

Hence, logicism as a paradigm of rationality leads to prescriptions, whereas logicism as a scientific paradigm guides explanations. However, the distinction between the two attitudes is not as sharp as in the case where logicism is an aprioristic scheme. In our approach the methodological rules draw their justification in part from their interaction with actual inferential practice. Moreover, logicism as a paradigm of rationality is fed back indirectly by inferential practice through its interaction with the methodological rules (as can be seen from the diagram).

The switch from one attitude to the other is therefore not so drastic. When we are in the justificatory mode of the system and we face a situation where the methodological rules are violated in many cases, we have the option to modify or replace our methodological theory; for instance, we may replace naive inductive theory by a Bayesian theory of confirmation, or a naive falsificationism by a sophisticated one (Lakatos, [1970]). If we cannot find a satisfactory logicist theory, however, we may look for another paradigm of rationality, but this weakens our normative stand significantly, for the following reason. In our search for a new paradigm, we aim at adopting a paradigm of rationality which will not clash too much with scientific practice. In other words, if we are willing to replace our paradigm of rationality in order to avoid the dilemma of the normative methodologist, we will adopt a paradigm which leads to a methodological theory which is not disconfirmed by scientific practice. In order to find out that this is the case, however, we must switch to the explanatory mode. In order to find out that a certain paradigm leads to a successful theory of science—*i.e.* a theory which does not clash too much with scientific practice—we must act as scientists rather than as normative methodologists. When such a paradigm is highly established throughout the theory it inspires, we might switch back to the justificatory mode and recommend that scientists employ the rules which are derived from our theory, knowing the limitations imposed on the normative strength of such recommendations.

In the above approach there is a smooth transition from justification to explanation and back since both have the same structure, *i.e.* the four-layered structure.

2 THE DISCOVERY-JUSTIFICATION DISTINCTION

2.1 *The Traditional D-J Distinction*

The discovery-justification (D-J) dichotomy is closely related to the descriptive-normative dichotomy. Hans Reichenbach introducted [1938] the distinction between the context of discovery, *i.e.* the actual processes leading to a new idea, and the context of justification, *i.e.* the way we validate the new idea. He maintained that the task of epistemology and philosophy of science is to deal only with the latter context, whereas the context of discovery is left to psychology, which can deal with the actual processes of thinking:

> Epistemology does not regard the processes of thinking in their actual occurrence; this task is entirely left to psychology. What epistemology intends is to construct thinking processes in a way in which they ought to occur if they are to be ranged in a consistent system; or to construct justifiable sets of operations which can be intercalated between the starting-point and the issue of thought-processes, replacing the real intermediate links. (Reichenbach [1938], p. 5)

Thus, epistemology is prescriptive ('ought') by virtue of its logical force, whereas psychology describes the actual ways we arrive at ideas. We may add also sociology, anthropology and even biology to the descriptive sciences which deal with discovery.

According to Harold Brown ([1977], p. 130), two theses are built into the D-J distinction: (a) that there is a sharp line separating the two contexts, and (b) that only the context of justification is amenable to logical analysis. We should add a third thesis implicit in the distinction: (c) that descriptive science (*e.g.* psychology) is irrelevant to the context of justification. Thus, the context of discovery can be dealt with by descriptive science, whereas the context of justification is normative. The proponents of this view maintain that the philosophy of science is normative and that it is a logic of science which should therefore deal only with the context of justification. An important consequence is that the context of discovery is irrelevant to the context of justification, since it is alogical.

The most widely cited illustration for the above theses is Kekule's discovery of the ring shape of the Benzen molecule while he was daydreaming about a snake chasing its tail. It would be implausible to claim that this process of discovery is amenable to logical analysis. Since only logically valid arguments can lead to justification or to epistemic warrant, the context of discovery here is irrelevant to the justification of Kekule's theory.

It should be stressed that the so-called context of justification encompasses all kinds of implications for the validity of a hypothesis. These include, in addition to confirmation and acceptance, also disconfirmation and rejection. Thus the term 'justification' is somewhat misleading. Indeed, Karl Popper, who rejects justification and accepts refutation only, joins in on the claim that the context of discovery is irrelevant to the logic of science. This is one of the cornerstones of his philosophy:

> The question how it happens that a new idea occurs to a man .... may be of great interest to empirical psychology; but it is irrelevant to the logical analysis of scientific knowledge. (Popper [1959], p. 31)

The term 'context of evaluation' may therefore be more appropriate than 'context of justification'.

Thus, according to the proponents of the D-J distinction, a philosophy of science is a logic of science. This is perhaps one of the major assumptions which shaped the 20th century philosophy of science, which had been dominated by logical empiricism and its offspring.

2.2 *Objections to the D-J Distinction and the Epistemic Dimension of Discovery*

2.2.1 The Distinction Fails in the Nonaprioristic Approach

The first objection to the above view arises when we adopt a nonaprioristic approach to justification, such as Goodman's. We recall that according to his analysis, the principles of logical inference themselves are justified if they are in reflective equilibrium with inferential practice. Hence, those principles which confer justification upon scientific claims, and which supposedly eliminate descriptive psychology from epistemology, are themselves justified by reference to actual thinking processes. Which discipline then, if not empirical psychology, can determine what are 'the particular deductive inferences we actually make', as referred to by Goodman? According to this analysis the epistemic status of a claim depends, among other things, on actual inferential practice, which is susceptible to psychological study. Hence, one of the implications of Goodman's analysis is that thesis (c) which is implicit in the D-J distinction fails, and with it the whole distinction. Indeed, if descriptive psychology infiltrates the context of justification, how will we distinguish between the two contexts? In this case neither of the two contexts would turn out to be purely logical. It follows from this that neither of the sides of the D-J dichotomy has a unique epistemic status.

A similar claim is raised by Hilary Putnam ([1975], p. 268): since the attempts to construct a logic of justification (such as inductive logic or probabilistic theories of confirmation) have failed, we may conclude that there is no valid algorithm of justification, but only maxims of justification, in scientific practice. Hence, the context of justification cannot be distinguished from the context of discovery, since both are guided by maxims; both contexts have the same epistemic status. Harvey Siegel ([1980] p. 301), however, raises an objection to Putnam's objection. He claims that

> the point of Reichenbach's distinction is that information relevant to the *generation* of a scientific idea is irrelevant to the *evaluation* of that idea; and this distinction between generation and evaluation (or discovery and justification) can be instructively maintained despite the fact that both contexts are guided by maxims.

Let us illustrate Siegel's claim by the following case. Suppose the construction of a theory is guided by maxims of adherence to a certain world view, or to some general theoretical principles. If, on the other hand, the maxims of evaluation depend only on the formal syntactic relations which hold between the theory and certain observational sentences, and not on the content of the theory, then justification does not depend on the maxims of discovery, which refer to the content of the theory. A simple example for such a situation is the case when the maxims of discovery demand that an explanatory theory of gas structure be corpuscularian and adhere to the mechanistic world picture, whereas the maxims of justification demand only that the logical implications of a theory match the observational results according to some formal confirmation theory. Here validation is indeed independent of the context of discovery.

Nevertheless, there is a serious flaw in Siegel's claim. By claiming that 'the point' of the D-J distinction is that the context of discovery is irrelevant to the context of justification, Siegel ignores the epistemological import of the distinction. According to Reichenbach and his followers, the two contexts differ in their epistemic status; the context of discovery does not confer an epistemic warrant to the discovered idea, whereas the context of justification does. Hence, if both contexts are maxim-guided, the epistemological spirit of the thesis will be absent even in the event that the context of discovery is irrelevant to the context of justification. The latter will have no normative import, *i.e.* it will not have any epistemological superiority over the former as the thesis requires. Hence the irrelevance of the context of discovery to the context of justification is a necessary but not a sufficient condition for the D-J distinction thesis. Putnam's objection refers, therefore, to the absence of epistemic superiority for the context of justification when it is maxim-guided.

2.2.2 Discovery Is Relevant to Evaluation in the Dynamicist View

The second objection to the thesis arises when we notice that there are important situations in which the context of discovery is essential to the context of justification, and consequently a necessary condition for the D-J distinction thesis is not met, and the thesis fails.

A very important question which bears on the confirmation of a scientific theory is whether or not a certain event or phenomenon which is predicted by the theory was known at the time the theory was proposed. In scientific practice it is well-known that when the predicted event occurs, or is discovered, only after the theory was proposed, then the theory's degree of credibility rises considerably, provided that the event has not been expected on other grounds. Typical examples are the discovery of a new planet (Neptune) or a new particle (the omega minus), which are predicted by physical theories. This methodological factor cannot be explicated in a confirmation theory which depends only on syntactical or formal relations between the theory and observation statements, since such relations are timeless, *i.e.* are insensitive to time priorities. Let us illustrate this point by the following example. Let us consider Nicod's rule of confirmation which states roughly that sentences of the form (∃x) (Ax & Bx) (*i.e.* 'there exists an object of the kind A with a property B') confirm the law-like statement (x) (Ax ⊃ Bx), *i.e.* 'every A is B'. This rule is insensitive to the question of when each statement became known to scientists. If our rules or maxims of justification are of the formal kind, like this one, then information as to how a claim has been arrived at is irrelevant to its confirmation. Typically, a formal rule of confirmation takes into account the final products of scientific discovery, *i.e.* the statements which describe laws and theories, without taking into account the history which led to their discovery. Thus the D-J distinction thesis fits in well with the ahistoric epistemological view, which maintains that all that is relevant to the validation of a claim is its formal structure and relations to other claims.

If we wish to stay close to actual inferential practice in science, however, we have to devise confirmation rules which will do justice to historical considerations such as the above one. With such rules, the question of whether or not a theory has been constructed with the knowledge of certain events will be relevant to the evaluation of the theory. The higher the percentage of facts previously unknown to the discoverer which is predicted by a theory, the higher will be the epistemic status conferred upon the theory by these rules. In such a case we can say that we have gained an epistemic profit from the theory. The theory is thus viewed from an epistemological point of view not as a finished product, but as a dynamic entity (see Kantorovich [1979]); epistemology has to assess it in its historical context. We may draw an analogy with testing a person with respect to his ability to solve a problem; it is not enough to consider his final solution since he may have gotten information concerning the solution before the test. The dynamicist view can be summarized by the following words of Frederick Suppe: 'Full epistemic understanding of scientific theories could only be had by seeing the dynamics of theory development' ([1974], p. 126). The D-J distinction therefore fails in the historicist view.

2.2.3 The Context of Discovery Has an Epistemic Dimension

If we assume that the search for truth is not blind, *i.e.* that among the infinite number of logically possible hypotheses humans frequently arrive, in particular in science, at hypotheses which later prove successful, it is unreasonable to exclude this fact from our epistemological discourse. We may require, therefore, that epistemology and philosophy of science account for this fact. According to this view science is not only an *evaluator* of ideas— whatever these may be—but more importantly a *generator* of successful ideas. Hence the ways by which humans, and in particular scientists, come to new ideas, are epistemologically important. There must exist a rational way to generate good ideas. We would otherwise be engaged in testing all kinds of hypotheses with no preferred direction, the chances for progress being accordingly diminished.

It may be instructive to draw an analogy with a sport such as running. The sport of running has two contexts, the context of training and the context of judging. According to the training-judging (T-J) distinction thesis there are no exact criteria as to how to produce a good athlete and to improve results. Only the measurement of the results achieved in a particular competition and the decision as to who is the winner are guided by exact criteria. Thus, the process of scientific discovery is analogous to the process of training. The product of the first process is a theory (or a law). The product of the second process is a trained athlete. The evaluation of a theory is analogous to the evaluation of the achievements of the athlete. The quality of a theory is judged with respect to its success in predicting empirical data and phenomena. The quality of an athlete is judged with respect to his scores in competitions. Furthermore, in the case of the sport of running, we cannot object to the claim that the methods of training and producing good athletes are irrelevant to the way of choosing the winner in a competition.

Hence, according to the T-J distinction thesis, a 'philosophy' or methodology of this sport should concentrate only on the methods of judging. This thesis would surely seem to offer too narrow a view of athletics, since it ignores the major goal of athletics, *i.e.* the goal of generating good athletes and breaking records. Moreover, if the aim of athletics is ever-improving achievement, it would be irrational for someone who wishes to understand this human activity to be solely engaged with judging, and not at all with the methods of improving achievement. It is of course entirely rational to be interested in methods of judging, but why call this a 'methodology of athletics'; a better name would be 'methodology of judging athletics'.

The moral in this for the philosophy of science is that it will miss the essence of science if it concentrates solely upon evaluation. Rationality in science resides not only in the activity of testing theories, but also, and perhaps mainly, in the activity of generating theories which are good candidates for testing. If science had been engaged only in testing theories, with no regard to which theories it was testing, it would have been in compliance with the D-J distinction thesis, but it is very doubtful that it would have arrived at its spectacular achievements in understanding and mastering natural phenomena.

Hence, the process of scientific discovery seems to be essential for the growth of scientific knowledge. With respect to this point, Popper says ([1969], p. 215) that 'continued growth is essential to the rational and empirical character of scientific knowledge'. Popper stresses that the way of growth as he conceives it—*i.e.* by conjectures and refutations—is responsible for the rational and empirical character of science. However, he maintains that it is not the business of (logical) epistemology to study the process of arriving at a new conjecture. But if the generation of conjectures were not guided by a mechanism of some epistemological merit, scientists would be engaged in criticizing slack hypotheses which lead nowhere. Hence, one of the tasks of explanatory epistemology should be to investigate this mechanism for generating conjectures.

A nonlogicist paradigm of rationality may have implications for the structure of scientific discovery. It may yield a theory of scientific knowledge which will explain both generation and evaluation of conjectures. Popper himself paved the way for such a theory, *i.e.* evolutionary epistemology, which was developed by Donald Campbell and his followers (see 'Evolutionary Epistemology Bibliography' in Callebaut & Pinxton [1987]). Evolutionary epistemology starts with the model of natural selection, which is based on blind variation. This model complies with Popper's view that theories are not constructed or inferred from observational data, or that theory-construction is logically blind to the data. However, different versions of evolutionary epistemology employ a variety of biological models for generating variations which have direct implications for the context of discovery. For example, a particular interpretation of the very model of blind variation, which employs the notion of serendipity, has essential implications for the structure of the process of discovery (see Ne'eman and Kantorovich [forthcoming]). This interpretation suggests that the most significant blind discoveries are made when scientists try to solve a given problem, or study a given phenomenon, but end up solving *unintentionally* another problem, or discovering a new phenomenon (*e.g.* Kepler, Planck, Roentgen and Fleming). This view of scientific discovery, which is related to certain biological models exhibiting stepwise evolutionary progress, leads to the conclusion that scientists do not gamble blindly with nature. Rather they should expect to make unexpected discoveries while being engaged in directed problem-solving within the framework of their current research programs. Evolutionary epistemology provides us, therefore, with examples of explanatory theories of scientific knowledge based on the evolutionary paradigm of rationality, which deal with both the contexts of discovery and evaluation and which are potentially prescriptive in the sense explained in Section 1.2.4.

The Institute for the History and Philosophy of Science and Ideas
Tel Aviv University

REFERENCES

Brown, H. [1977]: *Perception, Theory and Commitment.* University of Chicago Press.

Callebaut, W. and Pinxton, R. (*eds.*) [1987]: *Evolutionary Epistemology: a Multiparadigm Program.* Reidel.

Goodman, N. [1965]: *Fact, Fiction and Forecast.* Second edition. Bobbs-Merrill.

Haber, H. and Kane, G. [1986]: 'Is Nature Supersymmetric?', *Scientific American*, 254, p. 42.

Hesse, M. [1974]: *The Structure of Scientific Inference.* MacMillan.

KANTOROVICH, A. [1979]: 'Towards a Dynamic Methodology of Science'. *Erkenntnis*, **14**, pp. 251–73.

KANTOROVICH, A. [1982]: 'Quarks: An Active Look at Matter'. *Fundamenta Scientiae*, **3**, pp. 297–319.

LAKATOS, I. [1970]: 'Falsification, and the Methodology of Scientific Research Programmes', in I. Lakatos and A. Musgrave (*eds.*), *Criticism and the Growth of Knowledge*, pp. 91–196. Cambridge University Press.

LAKATOS, I. [1971]: 'History of Science and Its Rational Reconstructions', in R. C. Buck and R. S. Cohen (*eds.*), *Boston Studies in the Philosophy of Science*, **8**, pp. 91–136.

LAUDAN, L. [1977]: *Progress and Its Problems*. University of California Press.

NE'EMAN Y. and KANTOROVICH, A. [forthcoming]: *Science as Evolution and Transcendence*.

POPPER, K. [1959]: *The Logic of Scientific Discovery*. Sixth Impression 1972. Hutchinson.

POPPER, K. [1969]: 'Truth, Rationality and the Growth of Scientific Knowledge', in *Conjectures and Refutations*, pp. 215–50. Routledge and Kegan Paul.

PUTNAM, H. [1975]: *Mathematics, Matter and Method: Philosophical Papers* **1**. Cambridge University Press.

QUINE, W. V. [1969]: 'Epistemology Naturalized', in *Ontological Relativity and Other Essays*, pp. 69–90. Columbia University Press.

RAWLS, J. [1971]: *A Theory of Justice*. Harvard University Press.

REICHENBACH, H. [1938]: *Experience and Prediction*. University of Chicago Press.

SIEGEL, H. [1980]: 'Justification, Discovery and Naturalizing of Epistemology', *Philosophy of Science*, **47**, pp. 297–321.

STICH, S. P. and NISBETT, R. E. [1980]: 'Justification and the Psychology of Human Reasoning', *Philosophy of Science*, **47**, pp. 188–202.

SUPPE, F. (*ed.*) [1974]: *The Structure of Scientific Theories*. University of Illinois Press.

ZIMAN, J. [1968]: *Public Knowledge*. Cambridge University Press.

Brit. J. Phil. Sci. 39 (1988), 495–503 *Printed in Great Britain*

Statistical Explanation, Probability, and Counteracting Conditions

THOMAS R. GRIMES

---

1 *Introduction*
2 *Explanation by Degrees*
3 *Probabilities and Counteracting Conditions*
4 *Conclusions*

I INTRODUCTION

There is a host of troubling issues associated with the problem of statistical explanation. One issue concerns the extent to which such explanations are possible. Another focuses on the role of probability and whether it is necessary or even important for an adequate explanation to cite the probability of the explanandum event in relation to the explanans. And a third issue is whether counteracting conditions, conditions which lower the probability of the event to be explained, are explanatorily relevant and thereby to be included as part of a full and proper explanation. Though these issues are distinct, I will attempt to resolve the first issue and then show how the solution to this problem can be used to solve the other two.

2 EXPLANATION BY DEGREES

Three basic positions have been advanced regarding the proper scope and limits of statistical explanation. In defending what is perhaps the most popular view, Carl Hempel [1962], [1965] has maintained that it is possible to construct a sound statistical explanation, but only for those events capable of being predicted on the basis of the explanatory facts.[1] He writes:

> Any rationally acceptable answer to the question 'Why did event X occur?' must offer information which shows that X was to be expected—if not definitely . . . then at least with reasonable probability. Thus, the explanatory information must provide good grounds for believing that X did in fact occur; otherwise, that information would give us no adequate reason for saying: 'That explains it—that does show why X occurred.' ([1962], pp. 367–8)

---

[1] It should be pointed out that since his earlier influential writings, Hempel has changed his position on this issue and no longer insists on a predictability requirement. See the postcript to Hempel [1977].

For Hempel, then, there exists a fundamental asymmetry between events which are explanatorily probable, that is, probable in relation to the full set of relevant explanatory facts, and those which are not to the effect that the former but not the latter can be explained (*cf.* Nagel [1961], Rescher [1970], and Mellor [1976]).

Wesley Salmon [1984] finds Hempel's asymmetry thesis completely unacceptable. From an explanatory point of view, Salmon sees no essential difference between probable and improbable events, and since events which are explanatorily probable are explainable, likewise, so he argues, for events which are explanatorily improbable (*cf.* Jeffrey [1969], and Fetzer [1981]). That is,

> ... it appears that we understand the improbable outcome just as well as we understand the highly probable outcome. It strikes me as a rather peculiar prejudice which says that we can explain phenomena that occur frequently, but not those that occur infrequently. (Salmon [1984], p. 86)

To complicate matters, Georg von Wright [1971] in effect agrees with Salmon that the asymmetry thesis is untenable, but *contra* Salmon, von Wright maintains that both explanatorily probable and improbable events are altogether inexplicable, and thus challenges the very possibility of statistical explanation. On von Wright's view, the success of an adequate deductive-nomological explanation for a given event E resides in the fact that the explanation 'tells us why E *had* to be (occur), why E was *necessary* once the basis is there and the laws are accepted' ([1971], p. 13). This feature is necessarily missing, however, from an irreducibly statistical explanation, for as von Wright points out such an explanation 'admits the possibility that E might have *failed* to occur. It therefore leaves room for an additional quest for explanation: why did E, on this occasion, actually occur and why did it not fail to occur?' (*ibid.*). If this additional quest for explanation can not be fully satisfied, then according to von Wright there is no explanation for why E *occurred*. Instead, the most that can be said is that there is an explanation for why E was to be *expected*. For these reasons, then, von Wright rules out explaining any event whose occurrence was not fully determined on the basis of the explanatory facts (*cf.* Stegmüller [1973], and Harré and Madden [1975]).

Having outlined the basic features of these three competing views, I will argue for a fourth and compromising position which saves what I take to be the important insights of each of the other views.

As a first step in identifying the limits of statistical explanation, consider when an event may be said to be truly inexplicable. A paradigm case of an inexplicable event would seem to be one which occurred purely spontaneously and admits of no explanatorily relevant factors. For example, imagine a world where a certain strain of the flu develops spontaneously. In this world the probability of contracting the flu with respect to any prior condition C is the same as the probability of getting this illness with respect to not-C.[2] That is, Pr(F/C) = Pr(F/not-C), where F is the condition of having the flu; or alternatively, Pr(F/C) − Pr(F/not-C) = 0. In this world there is no explanation for why some individual, Jones, for example, happens to catch the flu. But now imagine a slightly different world where there is a single factor G, consisting in being exposed to certain germs, which is relevant to contracting the flu. In this other world G increases the probability of getting the flu, but suppose it is highly unlikely that this result will develop. That is, suppose that Pr(F/G) > Pr(F/not-G) = 0, but that Pr(F/G) = ·0001. Now if in this second world Smith should become exposed to the flu germs and as a result develop a case of the flu, could his illness be explained? Given that the probability of catching the flu once exposed to the flu germs is as low as ·0001, it seems that in an important sense we can *not* explain why Smith happened to catch the flu. After all, there is no whopping difference between his having the flu as a result of exposure to the germs, where Pr(F/G) − Pr(F/not-G) = ·0001, and the inexplicable spontaneous case involving Jones where Pr(F/G) − Pr(F/not-G) = 0. At the same time, however, Smith's coming to have the flu seems not totally inexplicable, for unlike Jones' situation we have at least *some* idea of why Smith became ill, to wit, he was exposed to the flu germs.

The proper way to reconcile these conflicting considerations, I would suggest, is to allow that Smith's condition *is* explainable but only *partially*. That is, if a particular event is inexplicable if its occurrence is purely spontaneous, then it seems not implausible to suppose that explanation, like a number of other important concepts, *admits of degree* and that an event whose occurrence is not fully determined on the basis of the explanatory facts, and hence is in part spontaneous, can not be fully explained. In the case of Smith, his being exposed to the germs is relevant to explaining why he caught the flu, but since the probability involved is only ·0001, it seems that Smith's contracting the flu was partly spontaneous and therefore only partly explainable. This assessment is particularly plausible if the relevant probability is construed, not as some relative frequency, but as a single case propensity, that is, where the expression 'Pr(A/B) = n' attributes to each particular state of affairs or event of type B a disposition of strength n to bring about a state of affairs of type A.[3] On this interpretation, there is an important sense in which Smith's bout with the flu would be less mysterious and better understood if the probability of catching the flu in relation to being exposed to the germs were, say, ·6 or 1·0 instead of ·0001. And the reason for this, it seems, is that in the former type of case, spontaneity has less of a role in the occurrence of the event and hence affords a better understanding of why it took place. Thus, Smith's case of the flu, though explainable, can only partially be explained.

---

[2] The negation here is to be understood in terms of narrow scope.

[3] See Fetzer [1981] for a single case propensity account of probability.

Some might object to this talk of an event's being only partially explained, but it is important to keep in mind the distinction between the explainability of an event and the explanation given of the event, where it is only the former that is being claimed to be partial in cases of statistical explanation involving indeterministic phenomena. An explanation, of any sort of event, may be regarded as complete if the explanation includes all the relevant explanatory information. However, if the occurrence of an event is not fully determined in relation to the explanatory facts, and hence is in part spontaneous, then the event itself is only partially explainable. In this way, an event may not be fully explained even though there is a complete explanation for why it occurred.

Allowing now that some events can only partially be explained, the issue arises as to how exactly this is to be measured. In the case of Smith considered above the matter is not too complicated, since in this situation there is only one explanatorily relevant factor, namely, being exposed to the flu germs, and this factor, though not sufficient, is necessary for coming down with the flu. Here the degree to which Smith's having the flu is explainable can be given the value of the conditional probability of his contracting the flu in relation to being exposed to the germs, or ·0001. Suppose though the situation is a bit different and being infected with the germs is neither necessary nor sufficient.[4] That is, suppose that although being exposed to the germs remains the only explanatorily relevant condition, coming down with the flu frequently occurs spontaneously. In this new situation, if Smith contracts the flu upon being infected with the germs where the probability of the former given the latter is, say, ·9—that is, $Pr(F/G) = \cdot 9$—yet having the flu can happen purely spontaneously such that $Pr\ (F/not\text{-}G) = \cdot 8$, the degree to which his bout with the flu is explainable can not be identified with the value suggested by the simple conditional probability of ·9 since this value fails to take into account the full force of spontaneity at work. Instead, the actual degree of explainability in this case is to be found by subtracting the element of spontaneity represented by $Pr(F/not\text{-}G)$ from $Pr(F/G)$ to yield a value of ·1. This difference properly discounts the effect of spontaneity and gives a true measure of the extent to which Smith's catching the flu can be explained.

This assessment assumes that there is only one explanatorily relevant factor to consider. Should there happen to be more than one such factor, the matter becomes somewhat more complex.

Suppose now that there are two conditions that are relevant to explaining a case of the flu, exposure to the flu germs and receiving a special type of cosmic radiation, condition H, where $Pr(F/G = \cdot 7$ and $Pr(F/H) = \cdot 6$. If the flu can occur spontaneously in the absence of both of these conditions such that $Pr(F/not\text{-}G\ \&\ not\text{-}H) = \cdot 2$, the degree to which Smith's having the flu given that he has been infected with the germs is not ·5, the difference between $Pr(F/G)$ and Pr(F/

---

[4] I am grateful to an anonymous referee of this journal for pointing out the importance of considering this type of case.

not-G & not-H), for the value expressed by Pr(F/G) may not rule out completely the influence of cosmic radiation. That is, the state of affairs consisting in being infected with the flu germs does not manifest as its only property the property of having the germs; it includes a whole host of other properties as well, one of which may be receiving the radiation. As a result, given the relevance of radiation, Pr(F/G) and Pr(F/G & not-H) may not express the same value; and since Smith happened to have the germs and not the radiation, it is the latter expression that we should be concerned with. So, to isolate the force of only that relevant factor that Smith actually possessed, the proper measure of the degree to which his having the flu can be explained is found by taking the difference between Pr(F/G & not-H) and Pr(F/not-G & not-H).

In general, then, if F is explained on the basis of G and G constitues a full set of the relevant explanatory factors, the degree to which F is explainable is the value expressed by Pr(F/G & not-H_1 &. . .& not-H_k) − Pr(F/not-G & not-H_1 &. . .& not-H_k), where H_1, . . .,H_k are nonobtaining factors that would have been relevant to explaining F had they obtained.[5]

If these remarks are correct and explanation is subject to degrees, there is then a simple compromise to be found in resolving the dispute between Hempel, Salmon and von Wright. Although von Wright is a bit extreme in ruling out the possibility of statistical explanation altogether, there is some merit to his claim that we can not explain an event which, given the explanatory facts, might have failed to occur, to wit, events of this sort are in part inexplicable and can not be explained completely. This idea of degrees also helps sustain Hempel's thesis that there exists an explanatory asymmetry between events which are explanatorily probable (that is, probable in relation to the explanatory facts) and those which are not. Where Hempel goes wrong, however, is in supposing that this asymmetry consists in our being able to explain the former but not the latter type of event. Instead, since events, for the most part, do not happen purely spontaneously in our world, the actual asymmetry (albeit a contingent one) is that explanatorily probable events are typically *more* explainable than explanatorily improbable ones. Indeed, in keeping with Hempel's concern that an adequate explanation afford grounds for expecting the event to be explained, both explanatorily probable and improbable events can be predicted; it is just that we have better grounds for expecting, and are more often correct about, those events which are probable and hence generally have a better understanding of why they occur. And finally, Salmon's proposal that both explanatorily probable and improbable events are capable of being explained, though essentially correct, stands in need of an important qualification to the effect that these events are in part inexplicable and can not be explained fully.

---

[5] In the anomalous case, not-G & not-H_1 &. . .& not-H_k may not be physically possible, though it is not clear whether this actually ever occurs. Nevertheless, this case can be handled by appealing to strong subjunctive considerations and identifying Pr(F/not-G & not-H_1 . . .& not-H_k) as the probability of F in the closest possible world where not-G & not-H_1 &. . .& not-H_k occurs.

By allowing that explanation admits of degree we see that statistical explanation is possible and applies to both explanatorily probable and improbable events—it is just that neither type of event is fully explainable and that what is explanatorily probable is generally more explainable than what is explanatorily improbable. This account of the limits of statistical explanation, it seems to me, is eminently plausible, especially given the way it establishes a compromise between the views of Hempel, Salmon, and von Wright. Further support for this position is gained by observing the ease with which it resolves the other two issues to be considered, the proper role of probability and the relevance of counteracting conditions.

3 PROBABILITIES AND COUNTERACTING CONDITIONS

It is typically held that an adequate statistical explanation should indicate the probability of the explanandum event in relation to the explanatory facts, but what is often missing is an account of why this piece of information is necessary or even relevant for the purposes of explanation.[6] Indeed, Paul Humphreys [1981], [1983] claims that identifying the overall probability is not important to the integrity of an explanation and that this value should in fact be omitted. On his view an adequate explanation need only specify some of the conditions that are causally related to the event to be explained.

Humphreys raises an interesting issue here. Once we have listed the various conditions explanatorily relevant to a given event, what purpose is served by going on to ascertain the probability of this event in relation to these conditions? After all, it is these conditions which do the explaining and once we have successfully listed them all, isn't adding more information in the form of an overall probability value simply excess baggage?[7]

The correct answer to all this seems to be that Humphreys is right and that reference to the overall probability need *not* be cited. This probability itself seems not to possess any special explanatory significance. However, once it is recognized that explanation admits of degree where the extent to which an event is explainable is measured as indicated in Section 2, the value obtained *does* provide a valuable piece of information that should be listed as part of the explanation. For example, if the value identified is rather low it is then apparent that there is much about the event's occurring that can not be explained; and if instead this value should happen to be quite high to the point of being 1·0, we then know that the event is fully explained. When this value is omitted, however, none of this information is available and we are left with simply a list of various conditions with no indication of the degree to which the explanandum event is explained. Thus, although it is not important that a scientific explanation specify the overall probability of the explanandum event in relation to the explanans, it is important that the degree to which the event is explainable be given.

---

[6] See, for example, Fetzer ([1981], pp. 130–1), who treats the overall probability value as a measure of the '*strength* of an explanans relative to its explanandum', but gives no account of why specifying this *strength*, which he takes to be essential, is important to the adequacy of an explanation.

[7] These are not the reasons Humphreys gives for requiring that the probability of the explanandum event not be cited. His concern seems to be that insisting on the correct probability value places an undue burden on an explanation by requiring that all the relevant conditions be specified and thereby making ordinary explanations exceedingly fragile.

The final issue to be addressed concerns the relevance of counteracting conditions. Intuitively an explanation for why a particular event occurred should include reference to those conditions which in some sense contributed to the event's taking place. This much seems clear, but suppose the event was not fully determined and that certain counteracting conditions were present which lowered the probability of the event by inhibiting or acting against its occurrence. Should these counteracting conditions also be cited as part of the explanation? As an illustration, if we wish to explain why Jones was tossed from his car while driving at the legal speed limit, it would be relevant to mention the fact that his car was struck by a cement truck; but is the fact that he was wearing a seat belt at the time, a counteracting condition in this case, also explanatorily relevant? Does specifying the fact that Jones had his seat belt fastened help explain *why* he got thrown from his car? Opinion on this matter is diverse. Both Hempel [1965] and Salmon [1984] require that any counteracting conditions be listed among the set of explanatory facts.[8] Bas van Fraassen [1980] and Raimo Tuomela [1981], on the other hand, develop theories requiring that these conditions *not* be cited.[9] And Humphreys [1981], [1983] takes the rather flexible view that counteracting conditions, though not entirely irrelevant, are not themselves explanatory and are to be cited only if some relevant noncounteracting conditions are mentioned. And even then, according to Humphreys, not every counteracting condition must be cited, for in his view a scientific explanation may be perfectly acceptable yet fail to identify all the relevant conditions, counteracting or otherwise. How is this tangled mess to be resolved?

Having argued that it is important for a complete explanation to identify the degree to which the explanandum event is explainable, it seems crucial that any counteracting conditions be cited as part of the explanation. If reference to these conditions is withheld, the value that results indicating the degree of explainability will be overinflated and thereby create a false sense of how much of the event's occurring is actually explained. That is, given that counteracting conditions have the effect of lowering the value given for the degree to which the event is explainable, to ignore these conditions would lead to underestimating the element of spontaneity underlying the event's occurring. Thus, in order to get a faithful and unbiased account of the degree to which the event is explainable, it is essential that the explanation cite whatever counteracting conditions were present.

---

[8] Salmon, unlike Hempel, is explicit about such a requirement. Nevertheless, for Hempel the relevance of counteracting conditions falls out of his maximal specificity requirement ([1965], pp. 399–401).

[9] This feature is not acknowledged by van Fraassen but is a consequence of his favoring criterion ([1980], pp. 47–8).

To anticipate a potential objection here, it might be claimed that it is implausible to require that counteracting conditions be mentioned as part of a fully adequate explanation if it is allowed that both explanatorily probable and improbable events can be explained, for this would have the untoward consequence that the very same explanation used to account for the occurrence of a given event E would also be used to account for its nonoccurrence, not-E, if E should fail to occur.[10] This objection, however, is largely unfounded. It is true that once counteracting conditions are included, E and its complement not-E will share the same set of explanatory facts (for any condition which helps bring about E will be a counteracting condition with respect to not-E, and vice versa), but it does not follow from this that E and not-E will have the same explanation. The degree of explainability associated with E typically will be inversely related to the degree associated with not-E, and since citing the value of this degree is properly a part of the explanation, this difference in value is sufficient to render the explanations for E and not-E distinct. Now in the rare case where both E and not-E feature the same degree of explainability, they would indeed share the same explanation, but even this situation should not be disturbing since it is clear that whichever event happens to occur will only be partially explainable. Thus, the objection here seems less than successful.

4 CONCLUSION

I have sought here to resolve three important issues associated with the problem of statistical explanation. I have argued that explanation admits of degree such that both explanatorily probable and improbable events can be explained but only partially, and have indicated how the degree to which an event is explainable is to be measured. Consequently, it is this value and not the simple probability of the explanandum event in relation to the explanans that is important to cite as part of a fully adequate explanation. This result in turn suggests that counteracting conditions are explanatorily relevant, for in order to determine the correct and objective degree to which an event is explainable, it is essential that any counteracting conditions be cited as part of the explanation.

Arkansas State University

---

[10] Meixner [1982] criticizes Salmon's statistical relevance model for having this feature.

REFERENCES

FETZER, J. [1981]: *Scientific Knowledge*. D. Reidel.

HARRÉ, R. and MADDEN, E. H. [1975]: *Causal Powers*. Basil Blackwell.

HEMPEL, C. G. [1962]: 'Deductive-Nomological vs. Statistical Explanation', in H. Feigl and G. Maxwell (*eds.*), *Minnesota Studies in the Philosophy of Science*, vol. 3, pp. 98–169. University of Minnesota Press.

HEMPEL, C. G.[1965]: *Aspects of Scientific Explanation*. Free Press.

HEMPEL, C. G. [1977]: *Aspekte wissenschaftlicher Erklärung*. Walter de Gruyter.

HUMPHREYS, P. [1981]: 'Aleatory Explanations', *Synthese*, **48**, pp. 225–32.

HUMPHREYS, P. [1983]: 'Aleatory Explanations Expanded', in P. Asquith and T. Nickels (*eds.*), PSA 1982, vol. 2., pp. 208–23. Philosophy of Science Association.

JEFFREY, R. [1969]: 'Statistical Explanation vs. Statistical Inference' in N. Rescher (*ed.*), *Essays in Honor of Carl G. Hempel*, pp. 19–28. D. Reidel.

MELLOR, H. D. [1976]: 'Probable Explanation', *Australasian Journal of Philosophy*, **54**, pp. 231–41.

MEIXNER, J. [1982]: 'Are Statistical Explanations Really Explanatory?', *Philosophical Studies*, **42**, pp. 201–7.

NAGEL, E. [1961]: *The Structure of Science*. Harcourt, Brace and World.

RESCHER, N. [1970]: *Scientific Explanation*. Free Press.

SALMON, W. C. [1984]: *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

STEGMÜLLER, W. [1973]: *Problems und Resultate der Wissenschaftstheorie und Analytischen Philosophie*. Springer-Verlag.

TUOMELA, R. [1981]: 'Inductive Explanation', *Synthese*, **48**, pp. 257–94.

VAN FRAASSEN, B. C. [1980]: *The Scientific Image*. Clarendon Press.

VON WRIGHT, G. [1971]: *Explanation and Understanding*. Cornell University Press.

Brit. J. Phil. Sci. 39 (1988), 505–519 *Printed in Great Britain*

Are All False Theories Equally False?

A Remark on David Miller's Problem and Geometric Conventionalism

THOMAS MORMANN

---

1 *Introduction*
2 *Quantitative False Theories*
3 *Nonisometric Isomorphisms of Metrical Spaces*
4 *Magnitude Spaces and Generalized Geometric Conventionalism*
5 *Topological* vs. *Metrical Structures*

---

1 INTRODUCTION

In a number of publications David Miller has argued (Miller [1975], [1984]) that we cannot make much sense of the idea that one false theory A is a better approximation to the true state of affairs T than is another false theory B. More precisely Miller shows by an elementary mathematical argument how, given any two independent physical magnitudes, f and g, on which theory A is the more accurate, we can always define in terms of f and g two new magnitudes h and k, for which the competing theory B gives the more accurate predictions. If the theory of special relativity is false that implies for example that we can specify a set of magnitudes for which any odd medieval astronomical theory (of course assumed to be false as well) will give better predictions than relativity mechanics. One may suggest that the existence of this set of magnitudes removes any grounds for regarding relativistic mechanics superior to any outdated medieval theory: neither theory will have a justified claim to being closer to the truth because we have no formal reason to prefer one set of magnitudes to the other one since they are inter-definable. If we had to accept Miller's result without qualifications we would have to conclude that the whole of empirical science is tottering, *cf.* (Popper [1979], p. 379). Of course this conclusion is unwarranted—rather than to deprecate empirical science we should tackle the philosophical task of improving our understanding about the real nature of empirical theories and their approximative relationships in order to avoid Miller's destructive scepticism. This philosophical problem is not trivial although Miller's argument is based on quite elementary mathematic's.

The aim of this note is to show that Miller's scepticism concerning the possibility of an objective ranking of false quantitative theories can be reformulated as a radical conventionalist thesis asserting that all metrical structures on a (higher than 1-dimensional) manifold of possible experimental outcomes of theories we A and B are on an equal footing, *i.e.*, it is a matter of convention which one choose just as it is a matter of convention if we drive on the left or on the right side of the road or use the unit 'meter' or 'yard' for measuring length.

Miller has considered his problem on different levels of sophistication. We shall deal only with the simple case when two competing theories A and B are engaged in the prediction of numerical *constants*. However, it should become evident in the following that the case of parameter depending magnitudes (*cf.* Miller [1975], [1984] and Horton [1978]) can be dealt with in a quite similar manner.

Before we start recall the following definition:

(1.0) Definition Let X be a nonempty set. A non-negative realvalued function $m: X \times X \rightarrow R$ is a metric on X if and only if the following conditions are satisfied for all $a,b,c \in X$:

(1) $m(a,b) = m(b,a)$
(2) $m(a,a) = 0$ and if $a \neq b$ then $m(a,b) \neq 0$
(3) $m(a,b) + m(b,c) \geq m(a,c)$

X provided with a metric m is called metrical space and denoted by (X,m). If there is no danger of confusion a metrical space (X,m) will be abbreviated by X.

We shall show that Miller's argument is based on the philosophical overestimation of the following well known geometric fact:

(1.1) Lemma Let E be a n-dimensional$_t$ linear space, $n \geq 2$. Let a,b and t be three mutually different points of E. Then we can define metrics m and m′ on E such that the following holds:

$$m(a,t) < m(b,t) \text{ and } m'(a,t) > m'(b,t)$$

This means that with respect to m the point a is a better approximation to t than b whereas with respect to m′ the point b is a better approximation to t than a.

The relevance of (1.1) for empirical matters should not be overestimated—consider the following simple example:

Suppose A and B are two rival theories about the performances of missiles predicting their impact under certain starting and flying conditions. To assess the respective approximative prediction power of A and B one has to define a metric m on the target field. Now suppose A has been worse than B with respect to m. That is we find the following distance relation between the predicted impacts a,b, and the true on t:

(1.2) $m(a,t) > m(b,t)$, a predicted by A, b predicted by B

If (1.1) could be applied the proponents of A could feel entitled to substitute m by some contrived metric m′ such that with respect to m′ A's performance is in better accordance with the true state of affairs t than B's with the argument that according to (1.1) one metric is as good as any other one.

For several reasons such a move seems hardly acceptable. We just mention two:

(i) because the *prediction* power of the competing theories is to be assessed the choice of a metrical structure m on the target field has to be made in advance and independently of the theories A and B.

(ii) even if—by lucky coincidence—the contrived metric m′ would have been chosen before the tests had been carried through (so that no question of prediction arises) nobody would accept m′ as a physically meaningful metric; of course, we have to explain the difference between a physically meaningful and a contrived metric. This topic will be discussed in some detail in Section 4. Thus we may say that even if it is formally possible to change the metrical structure of the target field there exist strong substantial arguments against this formal possibility. This issue will be discussed in more detail in the last two Sections of this paper.

After this disgression the proposed geometric reformulation of Miller's result can now be stated as follows:

(1.3) Let A and B be two competing theories engaged in the prediction of a pair of independent physical constants. Then the following holds:

(1) These constants have to be considered as the (predicted or true) values of certain magnitude variables f and g running over domains F and G respectively.

(2) A pair (f,g) of independent physical magnitudes defines a metrical structure on a linear 2-dimensional space Z of possible experimental outcomes of the rival theories A and B.

(3) Interdefinable pairs (f,g) and (h,k) define (usually nonequivalent) metrical structures $m_{(f,g)}$ and $m_{(h,k)}$ on a common space Z of possible experimental outcomes.

(4) If all metrics on Z have equal rights the ranking of the false theories A and B by approximative accuracy—depending on the metric chosen—is arbitrary.

We shall prove (1), (2), and (3) in the next two Sections. In Sections 4 and 5 we shall deal more explicitly with (4) in the context of geometric conventionalism pointing at a strategy how in this context Miller's scepticism may be resolved or at least defused.

2 QUANTITATIVE FALSE THEORIES

We begin the discussion of Miller's thesis with an example devised by Popper in the new appendix of Popper [1979], *cf.* also Miller [1984].

Let A and B be two rival theories engaged in predicting the values of two independent numerical constants f and g. Suppose that their predictions, along with the values given by the true theory (or by observation) T are as stated in the following table:

	f	g
A	·150	1·225
B	·100	1·000
T	·000	1·000

(2.1)

It seems obvious that B is a better approximation to T than A because its values both for f and g are in better accordance with the values given by T than those of A. Before we discuss Miller's argument against this *prima facie* plausible assertion let us make more explicit the *geometric* presuppositions underlying (2.1).

In order to interpret (2.1) as a table that gives us the values of some empirical constants, *e.g.* temperature and pressure constants of a certain gas, we must interpret f and g as the values of certain *variables* running over their respective domains F and G. That is to say, even if almost all elements of F and G are exluded by the theories A, B and T nevertheless the predicted and the true values make sense only if they are considered as elements of F and G. The statements '$f = \cdot 500$' or '$g = 5\cdot 000$' may be grossly false but they have to make sense if the predicted and the true values should have any meaning at all. Further, if the *true* values of f and g are obtained by using some measurement devices (as it is usually the case) strictly speaking we are not able to determine unique values but we can only localize some intervals due to the unavoidable inexactitude of measurement. This should suffice to make (1.3) (1) plausible.

Thus, if (2.1) is to be read as a table of the theoretically predicted and the true values of some empirical *constants* it is understood that these constants are values of magnitude variables running over some underlying magnitude spaces F and G. We therefore denote more precisely the value of f predicted by A by f_A, the meaning of g_B, g_T etc. defined analogously, of course $f_A \in F$, $g_B \in G$ etc.

In (2.1) it is assumed further that F and G are 1-dimensional *metrical* spaces such that the true and the predicted values of the constants in question can be characterized uniquely by their distances from certain gauge points $\cdot 000_F$ and $\cdot 000_G$ of F and G.

These are the geometric presuppositions we have to take into account. They may be denoted shortly as follows:

$$(F,m_F) \doteq (R,d) \text{ and } (G,m_G) \doteq (R,d) \qquad (2.2)$$

where m_F and m_G are metrics on F respectively G and (R,d) is the line of real numbers provided with its canonical metric.

The crucial premise of Miller's argument is that there are *two* (or more) independent magnitudes and that the predictive power of the competing theories A and B is measured with respect to *both*. In a natural way this task can be achieved using the *product space* $F \times G$ provided with the *product metric* $m_{F\times G}$ defined canomically as follows:

$$m_{F\times G}((x,y),(x',y')) = \sqrt{m_F(x,x')^2 + m_G(y,y')^2},\ x,x' \in F,\ y,y' \in G$$

Summarizing the above considerations we may say that the determination of the two numerical constants f_T and g_T (and of f_A,g_A etc. as well) should be conceptualized geometrically, *i.e.* as the localization of the point (f_T,g_T) in a 2-dimensional linear space $Z:=F\times G$ of possible experimental outcomes of the theories A, B and T. This provides (1.3) (2).

The space Z can be considered as a kind of configuration space or state space of the theories A, B and T. It may be noted that the concept of state space is one of the central notions of the so called semantic view of empirical theories; for a brief and non-technical review of this approach (*cf.* van Fraassen [1986]). However, it is not the purpose of this note to go into these matters in any detail, we shall use Z—the common state space or space of possible experimental outcomes of the theories A, B and T—solely as a convenient tool for the description of their approximative relationships.

Apparently the definition of Z depends on the specific pair of magnitudes (f,g) we have chosen for its definition. The next task is to arrive at a definition of Z which depends only on the *class* of interdefinable pairs of magnitudes and is independent of the specific pair we have chosen. This will be done in Section 3.

Using (2.3) the approximative prediction power of A and B can be determined by measuring the distance between (f_A,g_A) and (f_T,g_T), and (f_B,g_B) and (f_T,g_T). That is, we are inclined to judge the performance of A and B by the following criterion:

(2.4) The theory A is more accurate than the theory B (has greater approximate prediction power) if the following holds:

(*) $m_{F\times G}((f_A,g_A),(f_T,g_T)) < m_{F\times G}((f_B,g_B),(f_T,g_T))$

(*) is abbreviated by $m_{F\times G}(A,T) < m_{F\times G}(B,T)$.

Miller uses a somewhat different criterion:

(2.4)′ The theory A is more accurate than the theory B if and only if the following holds:

f_A is closer to f_T than f_B *and* g_A is closer to g_T than g_B.

(2.4)′ can be geometrically expressed as:

(2.4)″ The theory A is more accurate than the theory B if and only if the following holds:

(*)′ $m_f(f_A,f_T) < m_F(f_B,f_T)$ and $m_G(g_A,g_T) < m_G(g_B,g_T)$
(*)′ will be abbreviated $m_{F\times G}(A,T) <' m_{F\times G}(B,T)$

Obviously the following holds:

(2.4)‴ If $(m_{F\times G}(A,T) <' m_{F\times G}(B,T))$ then $(m_{F\times G}(A,T) < m_{F\times G}(B,T))$.

As Horton has pointed out (Horton [1978]) *cf.* also Miller [1975], a uniform superiority of one theory over the other is not the generic case, more typical than (2.1) are comparison tables of the following mixed kind where no theory has a monopoly of accuracy:

	f	g
A	·150	1·225
B	·151	1·000
T	·000	·000

(2.5)

It is intuitively plausible that in (2.5) the theory B is in better accordance with T than A. This intuition cannot be corroborated by (2.4)″ but by (2.4):

$$m_{F\times G}((f_A,g_A),(f_T,g_T)) = 1{\cdot}234 \quad m_{F\times G}((f_B,g_B),(f_T,g_T)) = 1{\cdot}011$$
$$\text{i.e. } m_{F\times G}(A,T) > m_{F\times G}(B,T) \qquad (2.6)$$

Therefore, if we consider (2.5) to be the generic case rather than (2.1) it is justified to use the criterion (2.4) instead of Miller's original (2.4)″ to measure the approximative prediction power of competing false theories. In any case Miller's argument against the plausibility of (2.4)″ holds for (2.4) as well without any change. That is if we use different but nonetheless equivalent magnitudes h and k instead of f and g the ranking of A and B according (2.4) is *reversed*.

We now describe this result in our geometric framework thereby proving (1.3) (3). Using (2.4) there is no essential difference between the tables (2.1) and (2.5) so we stick to the Popper-Miller example (2.1) throughout.

3 NONISOMETRIC ISOMORPHISMS OF METRICAL SPACES

Following Popper and Miller let us assume that we have a pair (h,k) of magnitudes interdefinable with the original pair (f,g) by the following equations:

$$\begin{array}{lcl} h:=g-2f & & f:=k-2h \\ & \text{and} & \\ k:=2g-3f & & g:=2k-3h \end{array} \tag{3.1}$$

With respect to h and k we get the following translation of the table (2.1):

	h	k
A	·925	2·000
B	·800	1·700
T	1·000	2·000

(3.2)

According to Section 2 these new magnitudes give rise to a metrical product space

$$(H\times K, m_{H\times K}) \doteq (R^2, d^2) \tag{3.3}$$

Using (2.4) we get

$$m_{H\times K}((h_A,k_A),(h_T,k_T)) = \cdot 075$$
$$m_{H\times K}((h_B,k_B),(h_T,k_T)) = \cdot 361 \tag{3.4}$$

i.e. $m_{H\times K}(A,T) < m_{H\times K}(B,T)$

In agreement with naive intuition now theory A seems to be superior to B. As Miller rightly emphasizes there is no formal reason to prefer either (f,g) or (h,k). We therefore seem to be involved in a remarkable dilemma because we obtain just the opposite ranking for the original pair (f,g):

$$m_{F\times G}((f_A,g_A),(f_T,g_T)) = 1\cdot 234 \quad m_{F\times G}((f_B,g_B),(f_T,g_T)) = 1\cdot 011 \tag{3.4$'$}$$

i.e. $m_{F\times G}(A,T) > m_{F\times G}(B,T)$

This situation may be summarized as follows: we have two different metrical magnitude spaces F × G and H × K giving rise to two incompatible rankings of the false theories A and B. Now, we have almost reached (1.1). We only need one further step:

(3.5) The equations (3.1) define in a natural and unique way linear maps r and s between H × K and F × G:

r: $H\times K \rightarrow F\times G$ by $r(h) := (-2f, g)$ and $r(k) := (-3f, 2g)$
s: $F\times G \rightarrow H\times K$ by $s(f) := (-2h, k)$ and $s(g) := (-3h, 2k)$

As can be seen immediately r and s are continuous and inverses of each other, *i.e.* $s.r = id_{H\times K}$ and $r.s = id_{F\times G}$, on the other hand r and s are *not* isometries, *i.e.* they do *not* preserve distance relations

$$m_{H\times K}(x,y) \neq m_{F\times G}(r(x),r(y)) \text{ and } m_{F\times G}(u.v) \neq m_{H\times K}(s(u),s(v)) \qquad (3.6)$$

Abbreviate $F\times G$ and $H\times K$ by X and Y. Define a space Z as follows:

$$Z := \{\{x,s(x)\},\ x\in X\} \text{ or equivalently } Z := \{\{y,r(y)\},\ y\in Y\}. \qquad (3.7)$$

Let m_X and m_Y be the metrics on X and Y defined above. They define metrics $\tilde{m}_X$ and $\tilde{m}_Y$ on Z in a canonical way:

$$\tilde{m}_X(\{x,s(x)\},\ \{x',s(x')\}) := m_X(x,x')$$

or equivalently

$$\begin{aligned} &\tilde{m}_X(\{y,r(y)\},\{y',r(y')\}) := m_X(r(y),r(y')) \qquad (3.8)\\ &\tilde{m}_Y(\{y,r(y)\},\{y',r(y')\}) := m_Y(y,y') \end{aligned}$$

or equivalently

$$\tilde{m}_Y(\{x,s(x)\},(x',s(x')\}) := m_Y(s(x),s(x'))$$

We have canonical isomorphisms $i: X \xrightarrow{\cong} Z$ and $j: Y \xrightarrow{\cong} Z$ defined as follows:

$$i(x) := \{x,s(x)\} \text{ and } j(y) := \{r(y),y\},\ x\in X,\ y\in Y \qquad (3.9)$$

It is evident that i is a $(m_X - \tilde{m}_X)$-isometry and j is a $(m_Y - \tilde{m}_Y)$-isometry. Thus the existence of two isomorphic but not necessarily isometric metrical spaces X and Y gives rise to the existence of two metrics $\tilde{m}_X$ and $\tilde{m}_Y$ on a space Z that is isometric to both:

$$(X,m_X) \cong (Z,\tilde{m}_X) \text{ and } (Y,m_Y) \cong (Z,\tilde{m}_Y) \qquad (3.10)$$

Remembering that m_X and m_Y are constructed with the aid of the magnitude pairs (f,g) resp. (h,k) the proof of (1.3) (3) is completed.

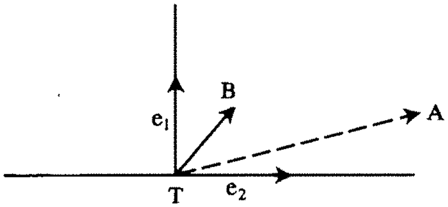
Thus we may reformulate (1.3) (4) in the language of (Grünbaum's) geometric conventionalism as follows:

(3.11) A n-dimensional linear space Z ($n \geq 2$) of possible experimental outcomes for some competing theories A and B is intrinsically metrically amorphous, *i.e.* Z has no 'built in' uniquely specified metric. Consequently, the ranking of the false theories A and B, as depending on the metrical distance relations to the true state of affairs T, is arbitrary.

4 MAGNITUDE SPACES AND GENERALIZED GEOMETRIC CONVENTIONALISM

It is not the aim of this note to go into the matters of geometric conventionalism in any greater depth (*cf.* for this topic, Grünbaum [1973], Putnam [1975], Diederich [1985] or Nerlich [1976]. For our purposes it is sufficient to describe a special kind of geometric conventionalism favoured especially by Grünbaum as the thesis that the metrical structure of physical space is a matter of convention, *i.e.* physical space is metrically amorphous and may be metricized in many different ways. We may characterize generalized geometric conventionalism as the thesis that higher dimensional physical magnitude spaces are metrically amorphous as well—this is just Miller's thesis (3.11).

Geometric conventionalism has been vigorously criticized by Putnam (Putnam [1975]). The aim of this Section is to sketch a similar line of criticism of Miller's generalized conventionalism. For this purpose consider once again our introductory example for rival theories A and B predicting the performance of missiles. For the sake of simplicity we may assume that the target field is a plane E provided with the usual euclidean metric m. m can be uniquely characterized by the stipulation of a pair (e_1,e_2) of orthonormal vectors. The prediction of the theories A and B and the true impact T of the missile may be located as follows:



(4.1)

(Without loss of generality T may be chosen as the origin of the coordinate system.) The vectors TA and TB are linearly independent, we may therefore define a new metric m* on E by stipulating that the following two vectors are orthonormal:

$$e^*_1 = 10TA \text{ and } e^*_2 = 1/10TB \tag{4.2}$$

e^*_1 and e^*_2 are new magnitudes clearly interdefinable with the original ones e_1 and e_2. We obtain the following comparison table for them:

	e^*_1	e^*_2
A	·100	·000
B	·000	10·000
T	·000	·000

(4.3)

Although no theory has a monopoly on accuracy with respect to e^*_1 and e^*_2 A clearly seems to be in better overall accordance with T. This is corroborated using the criterion (2.4):

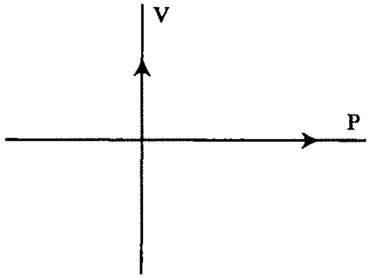
$$m^*(T,A) = \cdot 100 \quad \text{and} \quad m^*(T,B) = 10\cdot 000 \tag{4.4}$$

The proponents of A could indeed be happy with m*: according to it A seems 'a hundred times better' than B. For the sake of the argument let us further assume that by lucky coincidence m* were defined *before* the tests for A and B have been performed so that no question of *predictive* power arises. Even in this case nobody would assert seriously that A is 'really' better than B. Why? The answer is: m* is not a metric structure of physical space at all, it has no physical significance whatsoever. To argue in favour of this assertion we may follow a line of argument already used by Putnam (*cf.* Putnam [1975]).

The meaning of a term (in our case a physical metric or distance function) is *not* exhausted by a short list of axioms [*cf.* (1.0)] but is rather a function of an extended net of empirical knowledge. That is to say we do *not* fix the reference of the term 'metric of physical space' by convention but by *coherence*. The fixation by coherence involves large parts of scientific background knowledge and proceeds in a series of approximations. A first step for the fixation of a physically meaningful metric of physical space is to impose the condition that a measuring rod is to stay the same length when transported. Reichenbach erroneously thought that this condition would be sufficient to determine the metric of physical space uniquely but at least this condition excludes our contrived metric m*. Further steps of the approximation process may take into account constraints concerning the form of physical theories. *e.g.* invariance principles. It seems that in the case of physical space and its metrical structure this narrowing down process is quite successful: for middle sized objects and distances there is no other candidate left than the traditional euclidean metric of physical space. The more general case of arbitrary magnitude spaces has not been dealt with in greater depth. But there does not seem to exist a principle obstacle which would prevent us from adopting a similar approximation procedure in this generalized case too.

Consider the example of a magnitude space defined as the product space of two linear manifolds 'pressure' P and 'volume' V:

$p \in P,\ v \in V$



(4.5)

P and V are considered as orthogonal, suitably gauged they define independent magnitudes denoted by p and v and a metric on the 'pressure-volume' space $Z = V \times P$. Why do we not accept a new metric defined on Z with the aid of new magnitudes p_1 and v_1 in the following way:

$$\begin{aligned} p_1 &= ap + bv \\ & \qquad a,b,c,d \in R \qquad (4.6) \\ v_1 &= cp + dv \end{aligned}$$

To assure that (p,v) and (p_1,v_1) are interdefinable we have to assume that $ad - bc \neq 0$. Granted this would we accept p_1 and v_1 as serious physical magnitudes? I think not. Two arguments against the acceptance of p_1 and v_1 as honest physical magnitudes run as follows:

(1) The canonical metrical structure of the 'pressure-volume'-space Z is especially suited for representing properties of many empirical systems by rather simple geometric substructures of the state-space Z, *e.g.* curves of isotherms, vapor saturation curves of gases and liquids etc. A bad metric on Z could make the geometric representation of these systems extremely complicated quite analogously as a bad metric of physical space could make the representation of a simple transportation of a rigid rod utterly difficult. Generally new magnitudes p_1 and v_1 will not be able to compete in this regard with the classical magnitudes p and v and therefore would be rejected as physically meaningless. Thus the mere logical interdefinability of (p_1,v_1) with (p,v) is not sufficient to establish them as serious rivals of the traditional magnitudes. It should be noted that the rejection of (p_1,v_1) does not depend on the assertion that they cannot be interpreted as genuine 'pressure' and 'volume' but rather on the fault of having no genuine applications in the realm of elementary thermodynamical systems. Of course, things would look different if we could find an interesting class of applications for the new magnitudes p_1 and v_1 but this is not a matter for logic.

(2) There is another reason to reject the metrization of Z provided by p_1 and v_1: these terms *do not* appear as magnitudes of established theories while p and v *do*—namely in mechanics and geometry. A proponent of p_1 and v_1 would have to convince us that p_1 and v_1 are more appropriate magnitudes in mechanics, hydrodynamics and geometry, *i.e.* he would have to give arguments for the revision of all these theories underlying the physical meaning of the metrical structure of Z. In order to fulfill this task purely theoretical arguments will not suffice: up to now all our measurement devices and technical apparatuses are constructed for the traditional magnitudes p and v. A proponent of p_1 and v_1 would have to give good arguments for changing our practice of engineering and measurement before we would be prepared to change the geometry of the 'pressure-volume'-space Z.

The argument (2) is not to be understood in the dogmatic sense that a certain set of (traditional) magnitudes is beyond any doubt—under certain circumstances we may be forced or at least be inclined to revise our system of magnitudes thereby changing the ordering by accuracy of certain false theories. That this may be a serious possibility is indicated by the following example kindly brought to my attention by an anonymous referee of an earlier draft of this paper: Suppose the original parameters f and g stand respectively for p^2, the square of momentum, and d, the distance above the ground, of a body in free fall. Then we can define new magnitudes h and k as follows:

$$h=(1/2m)f+m\gamma g, \quad k=(1/2m)f-m\gamma g \tag{4.7}$$

m being the mass of the body and γ acceleration due to gravity. (h,k) and (f,g) are interdefinable since we calculate

$$f=m(h+k) \text{ and } g=(1/2m)\gamma(h-k) \tag{4.8}$$

Thus h and k give the basis for a new metric. It can be easily calculated that the new metric is *not* equivalent to the metric defined by f and g, and—this being the point of the argument—h and k are honest parameters appearing in established theories. They are just the Hamiltonian and the Lagrangian of a body in a gravitational field. Thus, there may well be competing (false) theories A and B whose ordering by accuracy is really doubtful because we have two incompatible physically meaningful metrics on their state-space! It should be emphasized that this example is an important step forward compared with the rather academic discussion of the possibility of defining an alternative metric on the 'pressure-volume'-space Z. But I think one has to go further: What is needed is a detailed logical reconstruction of full-fledged competing empirical theories. If this has been achieved then we can ask interesting questions like 'Is there an ordering by accuracy of Hamiltonian and Lagrangian mechanics or are they empirically equivalent?' etc. In my opinion it is not a small merit of Miller's problem to stress the necessity of dealing with adequate and full pictures of empirical theories. Taking into account the important role of state-spaces a semantic approach in the style of van Fraassen, Suppe or Sneed is perhaps the most promising one—but that is not the topic of this paper.

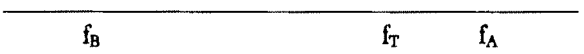
5 TOPOLOGICAL VS. METRICAL STRUCTURES

Miller has argued 'that any attempt to show some parameters to be more important, more significant, more fundamental, or whatever, than others is a symptom of an outdated—and fallacious—way of thought: essentialism. It is a way of thinking that takes words rather than theories far too seriously'. (Miller [1975a], p. 184.) This reproach may easily be returned: to hold the meaning of physical magnitudes depends solely on the satisfaction of some formal conditions (as they are laid down in our case in (1.0) takes logical formalities much too serious and may be characterized as a kind of *negative essentialism* that does not realize that many properties are part of a physical magnitude and that the meaning of such a term is not exhausted by a simple formal definition. The relevant structure of physical magnitude spaces is much richer and it depends in such complicated ways on other empirical theories and conceptions that it cannot bear just any prima facie possible formal manipulation as Miller asserts. This is made evident if we consider the different roles that Miller assigns to the topological and to the metrical structure of a magnitude space.

Even Miller does not accept just any formally possible manipulation of physical magnitudes. Arguing against Good (Good [1975], p. 205) who allows discontinuous transformations of (l-dimensional) magnitudes he rightly points out: 'Reversals of ordering by accuracy can of course be obtained even in the one-dimensional case if we are prepared to allow discontinuous transformations, . . . But this can scarcely be regarded as interesting, as some topological restraints must be insisted on if our reformulated hypotheses are to be thought of as reformulations at all.' (Miller [1984], p. 83.) To reveal the presuppositions of Miller's argument we make the following remarks:

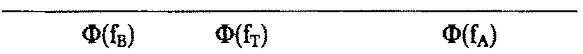
[1] Good's position is stronger than he himself and Miller realize, in many cases a reversal of ordering even for 1-dimensional magnitude spaces can be obtained by *continuous* transformations. This is seen as follows:

Let Z be a 1-dimensional magnitude space isometric to the real line R provided with its canonical metric. Suppose that the outcome of T, A, and B are located as follows:



(5.1)

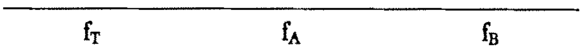
That is we have d(B,T) > d(A,T). But clearly there exists a homeomorphism (= continous map with continous inverse)Φ: $Z \xrightarrow{\cong} Z$ that switches the order of accuracy:



(5.2)

Of course Φ is not an isometry but it is at least continuous. If one considers Φ as inadmissible one is committed to the assertion that *not* all continuous transformations are allowed but only those that preserve the metrical structure of Z. But this means that one considers the metrical structure (and a fortiori the topological one) as *non-conventional.*

(2) Now we come to the case Good probably has had in mind as the only one. Suppose the distribution of the true and the predicted values of f is as follows:



(5.3)

If we are sticking to the canonical topological structure of Z it can easily be proved that for (5.3) an order reversing continuous transformation does not exist. But, if we are ready to change the topology of Z we can render continuous any Φ—take the new topology to be the discrete one—thereby obtaining a 'continous' order reversing transformation. 'Continuous' to be understood as continuous with respect to the new contrived topology. Clearly, Miller would not accept such a plot, his argument against Good rests on the assumption that the topological structure of Z is *not* changed, *i.e.* he takes the topological structure of the magnitude space as *non-conventional*. As has been explained in the preceding sections Miller considers the metrical structures of Z as *conventional* admitting maps that are continuous but do not preserve the metrical structure of Z (*cf.* Section 3). Thus the natural question arises: why is the metrical structure of Z conventional while its topological structure is not? As far as I can see Miller does not supply an answer to this question. I think that this is not his fault because presumably this question does not have a clear cut general answer. Rather, the only thing we can do is to examine carefully the physical meaning of the structure of the magnitude spaces under discussion in order to distinguish its conventional and non-conventional components at least partially. By no means is it obvious that the line between conventional and non-conventional structural ingredients can be drawn in all cases and once and for all such that the topological structure is non-conventional whereas all other stronger structures are conventional.

Taking the case of physical space as paradigmatic where an essentially unique metric exists one might conjecture that for many theories the class of physically meaningful metrical structures contains essentially one element that is to say all acceptable metrics can be transformed into each other by orthogonal transformations or by scale transformations. If this is the case the metrical structure of the state spaces of such theories is non-conventional, not all false theories are equally false, and Miller's recipe for the invention of new physical magnitudes would loose much of its philosophical appeal. However, it may be that our reliance on the uniqueness of the metrical structure of physical space is misleading especially in the more complicated cases of parameter depending magnitudes. In any case, this problem cannot be solved either by purely formal considerations or by mere intuitions. What is needed is a detailed structural description of the theory in question and its intertheoretical relations with other theories since the meaning of physical terms is not—so to speak—in the possession of insulated theories but is to be specified by greater parts of the net of empirical knowledge. In other words we have to take into account the holistic character of physical magnitudes. Thus, to counter successfully Miller's global scepticism concerning a stable ordering of false theories by approximative accuracy quite a detailed logical reconstruction of empirical theories will be necessary. Miller's result points at some rather deep issues of philosophy of science, despite its elementary mathematical character.

Freie Universität Berlin
Fachbereich Philosophie und
Sozialwissenschaften I

REFERENCES

DIEDERICH, W. [1985]: 'Conventionalism Reconsidered', in N. Rescher (*ed.*), *The Heritage of Logical Positivism*. Lanham/N.Y.: University of America Press, pp. 109–16.

GOOD, I. J. [1975]: 'Comments on David Miller', *Synthese*, 30, pp. 205–6.

GRÜNBAUM, A. [1973]: *Philosophical Problems of Space and Time*. 2nd enlarged edition, Boston Studies in the Philosophy of Science XII. Dordrecht: Reidel.

HORTON, R. J. [1978]: Accuracy of Prediction: a Note on David Miller's Problem, *British Journal for the Philosophy of Science*, 29, pp. 179–83.

MILLER, D. W. [1975]: 'The Accuracy of Predictions', *Synthese*, 30, pp. 159–91.

MILLER, D. W. [1975a]: 'The Accuracy of Predictions: a Reply', *Synthese*, 30, pp. 207–19.

MILLER, D. W. [1984]: 'Impartial Truth', in H. J. Skala, S. Termini, E. Trillas (*eds.*), *Aspects of Vagueness*, pp. 75–90. Dordrecht: Reidel.

NERLICH, G. [1976]: *The Shape of Space*. Cambridge: Cambridge University Press.

POPPER, K. [1976]: 'A Note on Verisimilitude', *British Journal for the Philosophy of Science*, 27, pp. 147–64.

POPPER, K. [1979]: *Objective Knowledge*, 2nd edition. Oxford: Oxford University Press.

PUTNAM, H. [1975]: 'The refutation of conventionalism', in H. Putnam, *Mind, Language and Reality*, vol. 2, pp. 153–191. Cambridge: Cambridge University Press.

VAN FRAASSEN, B. [1986]: 'Aim and Structure of Scientific Theories', in R. Barcan Marcus, G. J. W. Dorn, P. Weingartner (*eds.*), *Logic, Methodology and Philosophy of Science VII*, pp. 307–18. Amsterdam: North-Holland.

Brit. J. Phil. Sci. 39 (1988), 521–525 *Printed in Great Britain*

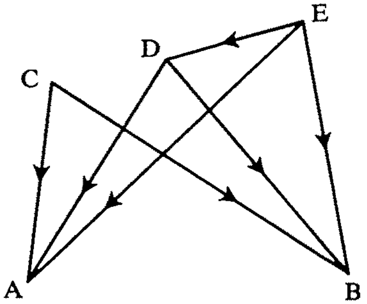
DISCUSSION

Response to Ehring's 'Papineau on Causal Asymmetry'

In 'Papineau on Causal Asymmetry' [1987] Douglas Ehring makes some telling objections to the theory elaborated in my 'Causal Asymmetry' [1985]. It seems to me, however, that his objections count rather more against the letter of my theory than its spirit.

My [1985] paper took off from the fact that 'screening off' is a pervasive feature of the natural world: probabilistic associations between joint effects disappear when we 'control for' (that is, conditionalize on the presence and absence of) their common causes. But in fact this was rather a loose way of describing what is really a more complicated phenomenon. The complication is that joint effects will often have a *number* of common ancestors in their causal history. And then their probabilistic association will only disappear if we control for *all* those common causal ancestors.

The kind of diagrams drawn in multiple regression and path analysis models of causal systems make the point clear.



To get the correlation between A and B to disappear, it's not enough to control for C. You also need to control for D and E. In general, you need to make sure that none of the lines coming into A are left with an uncontrolled-for common source with any of those going into B.

A number of writers, from Reichenbach on, have toyed with the idea that the screening-off asymmetry is constitutive of causal direction, that is, that the difference between causes and effects just *is* the fact that common causes screen off associations between joint effects but not vice versa. In 'Causal Asymmetry' I tried to go a step beyond this by explaining the screening-off asymmetry in terms of something else, and then making that something else constitutive of causal direction. The something else in question was the thesis that the various background conditions in conjunction with which causes fix their various effects are probabilistically independent of each other, while the 'background conditions' in conjunction with with effects 'fix' their various causes are not so independent.[1]

But now, if we consider diagrams like the above, we can see that this too is a rather loose way of putting something that really needs to be more carefully formulated. What we really need is something like the following thesis: in any *well-formed* causal diagram the various background conditions required for the tail of each arrow to fix the head of that arrow are probabilistically independent of each other, while the 'background conditions' required for the head of each arrow to 'fix' the tail of that arrow are not so independent (where what is required for a causal diagram to be 'well-formed' is that, if any two explicitly represented factors or relevant background conditions have a common causal ancestor, then that common causal ancestry should itself be explicitly represented in the diagram). And the overall theory then needs to be that *this* contrast is constitutive of the difference between causes and effects.

It might seem that the mention of common causal ancestry in the above explanation of what makes a diagram well-formed invalidates this theory as a reductive analysis of causal direction. But we can think of things in a way which eliminates this appeal to the concept of causal ancestry. Imagine that we start off with a map containing non-directional lines linking *all* factors which are causally connected to the phenomenon initially under consideration. Our task is to give some of these lines a causal direction. My thesis is that if we can find a way of doing this (in effect, a way of 'cutting out' a causal diagram) such that the background factors at the tail ends of the directional arrows are all robustly[2] probabilistically independent, then this is a well-formed causal diagram, and those arrows will be going from causes to effects, not from effects to causes.[3]

Once we have explained well-formedness in this way, we have a non-circular explication of the notion of causal direction. And then, given this entitlement to the notion of causal direction, we can go on to recognize that there will exist some diagrams whose arrows all correctly represent directed causal links, but which *don't* have probabilistic independence of background factors at the tail ends of the causal arrows: for a given set of causal arrows may merely give us a *sub*-diagram of a full well-formed diagram, where the full well-formed diagram would show that the sub-diagram had omitted certain lines of common causal ancestry.

---

[1] 'Background conditions' are here to be thought of as factors which aren't explicitly represented in causal diagrams like the above; they are the implicit 'error terms' whose values in particular cases determine whether or not the explicitly considered causes give rise to their effects.

[2] The significance of 'robustly' will be explained below.

[3] It is interesting to note that something like this analysis is implicit in the use of regression and path analysis to build causal models. Regression or path coefficients are only taken to be of causal significance in so far as the exogeneous 'error terms' in the regression equations are probabilistically independent of each other.

Unfortunately, these complications were only gestured at in my original paper. There is only the brief passage on pp. 280–1 (*cf.* Ehring's p. 82). This left the necessary elaboration of my thesis very much implicit. Moreover, this passage also carried the unfortunate suggestion that the only kind of case in which two background conditions will have a common causal ancestry, and so be probabilistically dependent, is where they are joint effects of a common cause—when there are, of course, other possible cases of dependence-inducing common causal ancestry, such as where one background condition is itself a cause of part of the other, or where one background condition is identical with part of the other.

It seems to me that if my overall thesis had been formulated so as to accommodate these complications, then most of Ehring's objections would cease to apply. Ehring's two putative counterexamples on p. 83 involve background conditions that are wholly or partially identical; the same is true in the indeterministic example he discusses at the foot of p. 87; and in the example at the top of p. 84 he postulates pairs of background conditions where part of one condition causes part of the other. All of these are cases where the background conditions have a common causal ancestry (even though they aren't joint effects of a common cause), and so their lack of probabilistic independence wouldn't impugn a more careful formulation of my thesis.

The case Ehring discusses later on p. 84 is slightly more complicated. Here he aims to show that it is perfectly possible to find probabilistic *independence* among the 'inverted background conditions' in conjunction with which common *effects* fix their joint causes, contrary to my claim that such independence is what distinguishes causes from effects. However, he achieves this by supposing that one of the 'background conditions' in question (R) is always either caused by another background condition (A), or by a factor (B) which is negatively correlated with A. In general, such a negative correlation indicates some kind of common prior ancestry; and of course A causing R is itself an instance of common causal ancestry. Once more, I would argue that when these common causal ancestries are taken into account, then the required probabilistic dependence of 'inverted background factors' will reappear.

This example also raises another difficulty, however. Even if the desired probabilistic dependence of 'inverted background conditions' can be restored by taking into account common causal ancestry, it remains the case that, prior to taking that common causal ancestry into account, we have an causal diagram with the background conditions at the tail end of the arrows probabilistically independent, even though the arrows in fact go from a common effect to its joint causes. So why isn't this a well-formed causal diagram, given my earlier remarks, and why doesn't it then refute my claim that the probabilistically independent tails of arrows in well-formed diagrams are always causes? Here I appeal to the requirement of *robust* probabilistic independence to which I drew attention in footnote 2. It is true that A and R appear probabilistically independent if we don't control for any further factors. But in practice, I claim, their 'real' probabilistic dependence would reappear once we did make explicit the relevant common causal ancestry. The background factors in a well-formed causal diagram, by contrast, are required to be robustly independent, in the sense that they have to remain probabilistically independent whatever further factors we being into explicit account.

Actually, whether the requisite probabilistic dependence between A and R would reappear in *Ehring's* example, as opposed to in a real case of similar structure, is a moot point: Ehring places such strict requirements on the correlations between A, R, and B that some of the relevant conditional probabilities become undefined. But I would argue that this makes Ehring's case unrealistic. In practical cases there always remain 'error terms' that have not been brought into explicit account, and the resulting 'unexplained variance' keeps the relevant conditional probabilities defined.

This raises a general point about my theory and Ehring's criticisms of it. Clearly, my theory isn't at all plausible as conceptual analysis: even before we go into details, it's obvious that there's nothing about probabilities or background conditions in our everyday concept of causal direction. However, my aim isn't conceptual analysis, but metaphysical reduction: I'm trying to identify the essential characteristics of the direction we intuitively recognize as the cause-effect direction. So examples designed to show that it's conceptually possible that a cause-effect relationship might lack those characteristics will cut no ice against my theory. I don't claim that such cases are conceptually impossible. My claim is simply that my theory identifies the real nature of the cause-effect relationship.

This last point is relevant to the very first objection Ehring brings to my theory (top of p. 83). Here he postulates a pair of joint effects of a common cause, whose respective background causal factors themselves have a common cause, and so on *ad infinitum*. Given this set-up, it would be impossible ever to take enough factors into explicit account to leave the remaining 'error terms' probabilistically independent, and so impossible to construct the kind of 'well-formed diagram' required for my theory to be applicable. Ehring's thought is that this clearly wouldn't stop the causes in the postulated set-up from *being* causes; and so causal direction must be something different from what my theory says. My response is to concede, for the sake of the argument, that such infinitely receding structures of common causes might be conceivable, but to deny that such structures can ever really occur.

My theory is based on the assumption that the screening-off asymmetry is the key to the nature of causal direction. The screening-off asymmetry is that associations between joint effects are screened off by their common causal ancestry, but associations between joint causes aren't screened off by their common descendants. Ehring's postulated case has joint effects whose association can't be screened off by controlling for their common ancestry. Maybe there is a sense in which we can *imagine* such a case. But if I am right about the significance of the screening-off asymmetry, then the case we are imagining is one which lacks the structure on which causal asymmetry in fact rests, and so isn't really possible.[4]

D. PAPINEAU
Cambridge, U.K.

---

[4] I haven't discussed Ehring's final argument (p. 86), which is a consequence of some sloppy terminology on p. 284 of my original article. Ehring notes an ambiguity in my saying that some factor is required 'to ensure the cancer did come from the smoking and not from the asbestos'; and he points out that the surrounding argument doesn't make much sense on either reading. But I'm afraid I had in mind a yet third reading: I meant that it was only when conjoined with the factor in question that the cancer became a sufficient condition for smoking. I still think that the argument is all right on this reading.

REFERENCES

Ehring, Douglas [1987]: 'Papineau on Causal Asymmetry', *British Journal for the Philosophy of Science*, 38, pp. 81–7.

Papineau, David [1985]: 'Causal Asymmetry', *British Journal for the Philosophy of Science*, 36, pp. 273–89.

Brit. J. Phil. Sci. 39 (1988), 527–529 *Printed in Great Britain*

DISCUSSION

How Nancy Cartwright Tells the Truth

In a recent paper, Kline and Matheson [1986] have attempted to save the covering law model of explanation from Nancy Cartwright's extended criticism [1983]. Briefly, their defense consists of replacing the truth of the covering law providing the explanation with a notion that the law must be at least approximately true. In this note I will show that this defense is inadequate and the examples will argue that Cartwright's insistence that causal factors must be included in any adequate model of explanation is correct.[1]

I will begin by criticizing the notion of approximate truth. It is well known that this idea does not work when one compares two theories of the same phenomena which have different ontologies (Laudan, [1984]). In what way can a particle theory of light be said to be an approximation to the wave theory that succeeded it? Particles and waves are different and exclusive, although not exhaustive, descriptions of light. Except at the empirical level, where the particle theory was deemed to be adequate, there was no approximation to the ontology of the wave theory. This, in fact, seems to be what Kline and Matheson have in mind, namely that approximate truth involves empirical or numerical accuracy. Although they never provide a detailed description of what they mean, the example they use to illustrate it emphasizes such numerical accuracy.

> Let us introduce the model by means of an example. We could explain the trajectory of a baseball thrown from a tower as follows:
>
> (1) The distance of the tower from the surface of the earth is θ meters.
> (2) Newton's Law of Gravitation is $F=G^*m_1{}^*m_2/r^2$.
> (3) The downward acceleration due to gravity is G^*m_1/r^2.
> (4) There are no other forces present that would produce an acceleration of magnitude f, where f is very small.
> (5) Therefore the trajectory of the body will be d(t) plus or minus e(t) where e is the largest possible error due to perturbing forces. Kline and Matheson ([1986], p. 35).

The emphasis on numerical accuracy as approximate truth is made clear by (4) and (5).

The inadequacy of this notion of approximate truth is made clear by a simple example. Suppose we have a theory, T_1, in agreement with our present beliefs, that predicts the conservation of energy. This means that for a given system the difference in energy before and after any particular process will be identically zero. Let us consider another theory, T_2, that predicts that energy is not conserved and that for a particular process the difference in energy before and after will be 10 energy units. The energy difference is measured and found to be 2 energy units. It would seem that on any reasonable view T_2 is closer to the truth even though T_1 is numerically more accurate. The fact of energy nonconservation is far more important than the particular numerical value of the violation.[2]

---

[1] In this paper I am not defending all of Cartwright's views, but rather defending her against what I consider unjust criticism.

Further support for the inadequacy of this view of approximate truth as numerical accuracy comes from consideration of an example discussed by both Cartwright and by Kline and Matheson, that of Newton's and Kepler's laws and the motion of the planets. Using Newton's laws of motion and his inverse square law of universal gravitation Kepler's laws are strictly false because of the perturbations caused by the other planets. Let us take the advance of the perihelion of Mercury as a numerical measure of the accuracy of Newton's laws in this case. The deviation of the orbit of Mercury from a closed ellipse is approximately 527″ of arc per century.[3] Yet almost everyone will agree[4] that Newton's laws explain both Kepler's laws and the motion of the planets. One could say, and this is true, that Newton derived his results in an approximately and idealized way, (Franklin and Howson [1985] in agreement with the view of Kline and Matheson), and that the discrepancy is small. This seems plausible, but let us consider the situation in the late 19th century. Le Verrier showed that the Newtonian calculations, including the perturbations due to other planets, deviated from the measured value of the advance of the perihelion of Mercury by 39″ of arc per century.[5] In this case the deviation was regarded as arguing against the Newtonian explanation and various alternative explanations were considered.[6] Why was this, when the deviation was more than an order of magnitude smaller than the original discrepancy due to the perturbations of other planets? Clearly, numerical accuracy is not the answer. I suggest that it was because the cause of the original deviation from Kepler's laws, the perturbations due to the other planets, was known, even if the exact calculations could not be performed at the time of Newton. In the case of Le Verrier's deviation in the perihelion advance no cause of the discrepancy was known at the time. The failure of the alternative explanations, *i.e.*, the planet Vulcan, intra-Mercurial gas clouds, deviations from the inverse square law, velocity dependent forces, etc., only made the anomaly more severe.

---

[2] A Bayesian analysis can help to clarify this. Let us separate our experimental result into two parts e_1, the nonconservation of energy, and e_2, the numerical value of that violation. On the basis of our background knowledge, energy is conserved and so the probability of e_1 given any theory other than T_2 is very small. Using Bayes' Theorem:

$P(T_2|e_1) = P(e_1|T_2)\,P(T_2)/P(e_1)$

But $P(e_1|T_2) = 1$ because because $T_2 \vdash e_1$, and

$P(e_1) = P(e_1|T_2)P(T_2) + P(e_1|\neg T_2)P(\neg T_2)$

$P(e_1|\neg T_2)P(\neg T_2) = \epsilon$, where ϵ is very small.

Thus, $P(T_2|e_1) = P(T_2)/[P(T_2)+\epsilon] \approx 1-\beta$,

where $\beta = \epsilon/P(T_2)$, is also very small.

Set $P(T_2|e_1) = P_1(T_2)$, the new prior probability of T_2. Let us now consider what happens to our belief in T_2 when e_2, the wrong numerical value of the violation is measured.

$P_1(T_2|e_2) = P_1(e_2|T_2)P_1(T_2)/P_1(e_2)$

$P_1(e_2) = P_1(e_2\ T_2)P_1(T_2) + P_1(e_2|\neg T_2)P_1(\neg T_2)$

Let $K = P_1(e_2|\neg T)/P_1(e_2|T_2)$; then from above

$P_1(e_2) = P_1(e_2|T_2)P_1(T_2) + KP_1(e_2|T_2)[1-P_1(T_2)]$

but $P_1(T_2) = 1-\beta$,

$P_1(e_2) = P_1(e_2|T_2)\,[1-\beta] + KP_1(e_2|T_2)\,[1-1+\beta]$

$P_1(e_2) = P_1(e_2|T_2)\,[1-\beta+K\beta]$

Thus, $P_1(T_2|e_2) = P_1(e_2|T_2)P_1(T_2)/P_1(e_2|T_2)\,[1-\beta+K\beta]$

We expect $K<1$, because $\neg T_2$ is a weak statement compared with T_2, and in this case we see $P_1(T_2|e_2) \approx P_1(T_2)$, or our confidence in T_2 remains virtually unchanged. I do not suggest that any scientist actually engages in this two part process. It is done all at once. But I think the point is clear as to the relative importance of the violation of the conservation law and the numerical value of that violation. The violation of the conservation law is far more important.

[3] This numerical result was not known to Newton, although the effect was known. The numerical result was known in the 19th century.

[4] See any introductory physics text, for example Tipler ([1982], p. 349).

Cartwright's insistence on causal stories as part of an explanation seems correct. As she says, 'The emphasis on getting the causal story right is new for philosophers of science; and our old theories of explanation are not well-adapted to the job. We need a theory of explanation which shows the relationship between causal processes and the fundamental laws we use to study them and neither my simulacrum account nor the traditional covering law account are of much help.' Cartwright ([1983] p. 162).

ALLAN FRANKLIN
University of Colorado at Boulder

---

[5] This was actually the value obtained by Le Verrier. The modern value is 43″ of arc per century.

[6] For a detailed history of this episode see Roseveare [1982].

REFERENCES

CARTWRIGHT, N. [1983]: *How the Laws of Physics Lie*. Oxford: Clarendon Press.

FRANKLIN, A. and HOWSON, C. [1985]: 'Newton and Kepler, A Bayesian Approach', *Studies in History and Philosophy of Science*, 16, pp. 379–85.

KLINE, A. D. and MATHESON, C. A. [1986]: 'How the Laws of Physics Don't Even Fib', *PSA 1986*, Volume 1, pp. 33–41.

LAUDAN, L. [1984]: 'A Confutation of Convergent Realism', in J. Leplin (ed.), *Scientific Realism*, Berkeley: University of California Press.

ROSEVEARE, N. T. [1982]: *Mercury's Perihelion from Le Verrier to Einstein*. Oxford: Clarendon Press.

TIPLER, P. [1982]: *Physics*. New York: Worth.

Brit. J. Phil. Sci. 38 (1988), 531–535 *Printed in Great Britain*

DISCUSSION

A Green Parrot is Just as Much a Red Herring as a White Shoe: A Note on Confirmation, Background Knowledge and the Logico-Probabilistic Approach

In a recent paper Lawson [1985] has argued that the empiricist view of confirmation due to Keynes *et al* can avoid the famous paradoxes of confirmation through the explicit introduction of background knowledge and that this then removes the only advantage the so-called 'historical' account of Watkins *et al* has over the former. The attempt is based on the introduction of a 'specific field of application' in terms of a set F such that a generalization of the form 'all A's are B' from the perspective of existing background knowledge, can be expressed thus:

$$(\forall x \in F)\,[Ax \supset Bx]$$

with the logically equivalent contra-positive:

$$(\forall x \in F)\,[\sim Bx \supset \sim Ax]$$

In the case of the well-known ornothological version of the paradox, in which the generalization concerned is the statement 'all ravens are black', the only possibilities for F are that it is identical to the set of ravens or to the set of all birds and in either case such objects as white shoes are excluded as potential confirmers. Lawson then concludes that '(a)ny apparent paradox is thus avoided' (*Ibid.*, p. 399).

The purpose of this note is to point out a few elementary objections against such an attempt.

The first is that although the above manoeuvre may indeed avoid the most blatant and striking form of the problem, it in fact reappears on a more restricted level, within the 'specific field of application'. The counter-intuitive nature of Hempel's paradox essentially arises, as is well-known, because:

$$\sim Aa \cdot \sim Ba$$

offers a confirming instance of:

$$(\forall x)[\sim Bx \supset \sim Ax]$$

which is logically equivalent to:

$$(\forall x)[Ax \supset Bx]$$

I do not see how Lawson has shown that a similar sort of result does not arise in his account since even with $a \in F$ we still have:

$$\sim Aa \cdot \sim Ba$$

confirming

$$(\forall x \in F)[\sim Bx \supset \sim Ax]$$

and therefore also:

$$(\forall x \in F)[Ax \supset Bx]$$

In other words, staying within the class of all birds, the account does not avoid the counter-intuitive result that the observation of a green parrot confirms the hypothesis 'all ravens are black' just as much as the observation of a black raven.

Obviously the introduction of some such notion as a specific field of potential confirmers does have an intuitive plausibility, but this is really not the point. The problem is not just to show that white shoes do not have equal status with black ravens as potential confirmers of the hypothesis, but that neither do green parrots. What is required is some account of the relationship between hypotheses and evidence which will allow us to distinguish, from within a given 'field of application' if you like, the *extent* to which various pieces of evidence support a hypothesis.

Of course it might be said in reply that this is exactly where a qualitative confirmation function fails to deliver the goods and that what is needed instead is some quantitative account. This is the context in which Lawson claims to be operating, although most of the discussion is, in fact, couched in purely qualitative terms, with the single, rather tenuous, connection with the Keynesian probabilistic model provided by the flat statement that the former is consistent with the latter. (The relationship between the 'field of potential application' F and background knowledge k, for example, is nowhere made explicit beyond the remark that k 'influences' F, (*Ibid.*, p. 397), and is certainly not obvious.)

This brings us to our second criticism of Lawson's attempt, which is broadly to the effect that he fails to explain how the logico-probabilistic position which he advocates can avoid all the well-known objections usually levelled against its Bayesian counterpart. Furthermore, it appears that a typical Bayesian solution to the most famous of these actually runs counter to the general thrust of the Keynesian programme as a whole.

The first objection concerns the assignment of the initial probabilities p(h/k) fed into the usual belief change mechanism:

$$p(h/k) = \frac{p(e/hk)}{p(e/k)} p(h/k)$$

which expresses the degree of support given to hypothesis h by evidence e given background information k.

Lawson says only that the initial probabilities are obtained by analogy (*Ibid.*, p. 401). However, the Keynesian line which he adopts (*Ibid.*, p. 403), seems to be similar to the Bayesian defence that just as deductive logic is now recognized to be a purely conditional discipline, relating the truth values of certain sentences to one another, so, analogously, the inductive probabilistic view simply relates the rational belief-probabilities of certain statements to those of certain others (see, for example, Howson [1985], p. 306).

There are two points which can be made against this kind of response. The first is to deny the strength of the analogy and to argue, as Glymour does, that '. . . so far as understanding scientific reasoning goes, I think it is very wrong to consider our situation to be analogous to that of post-Fregean logicians, our subject matter transformed from a hodge-podge of principles by a powerful theory whose outlines are clear. We flatter ourselves that we possess even the hodge-podge' (Glymour [1980], p. 68).

The second point is that although it may be admitted that such accounts offer some insight into the structural relations between evidence and hypotheses, the question as to the coherency of the general principles restricting the assignment of the initial probabilities remains. Without such principles, of course, this view collapses into a serious form of subjectivism, despite Lawson's comments to the contrary (Lawson, *op. cit.*, p. 403).

The second, and very famous, problem is more serious as far as Lawson's attempt is concerned in that the Bayesian answer to it conflicts with the whole Keynesian approach: we suppose that $h \vdash e$ and that e is in fact part of the background knowledge k. In this case, p(e/hk) = 1 and p(e/k) = 1 and we have:

$$p(h/ek) = p(h/k)$$

Thus it appears that known evidence cannot increase the degree of confirmation of a hypothesis, a result which, in its generality, seems strongly counter-intuitive and contrary to what we know about the history of science—previously known and very old observations were taken as supporting Copernicus's theory, as were diffraction phenomena for Fresnel's wave hypothesis, the anomalous advance of the perihelion of Mercury for Einstein's field equations etc. etc.

A recent attempt to resolve this problem from the Bayesian standpoint (Howson [1984]), has argued that the probabilities should in fact be relativised to k-e rather than to the entire k, which obviously then implies $p(e/k) \neq 1$, and that this is entirely in tune with the general Bayesian line since '... the Bayesian assesses support by how your odds on h *would* change were you *now* counterfactually to come to know e that—in the circumstances envisaged—you already do know e means that you relativise to K-(e)' (Howson [1985], p. 307).

However this kind of solution is simply not available to the Keynesian who gives an account in which '.... background information is treated much like any other kind of evidence—the relation between it and relevant hypothesis being "purely logical"—...' (Lawson *op. cit.*, p. 403). The abandonment of the 'Principle of Total Evidence' and the separation of known evidence from background information cannot be tolerated from this point of view, since no distinction is made as regards the temporal character of the evidence. If such a treatment were to be admitted then of course Lawson's argument would lose a lot of its point since the aim was to emphasise the advantages of this position over the 'historical' approach which is entirely based on such a distinction.

It is prediction, of course, that occupies an important role in the latter account and old evidence, even when viewed counter-factually as above, is not predicted evidence. Nevertheless, adopting the Bayesian answer to the above problem would clearly remove the force of the claim that there is no aspect of the Keynesian account '... that depends on the historical ordering in which e and h were obtained' (*Ibid*, p. 401). A failure to adopt some such account would be even worse, of course.

Finally there is the question of the lack of historical support for these probabilistic views of confirmation. This is a well-known objection in the context of the physical sciences, whether it exists for economics also, where Lawson is concerned to reintroduce the logico-probabilistic account, is not discussed. Indeed, Lawson simply states quite bluntly that this account is '... in accordance with research practice as we observe it' (*Ibid.*, p. 400), without giving so much as a shred of evidence for this claim.

This is quite separate from the question whether the formulation of hypotheses ever involves the statement of the potential field of application, although Lawson himself seems to conflate the two. In this case I am inclined to agree with him against Hempel that such considerations do enter into scientific reasoning, at least implicitly, but clearly further discussion is needed here, in particular as regards such problems as the widening of the field of application through inductive generalizations, such as occurs in Newton's argument for universal gravitation for example, and the unification of theories where the fields of application are different, to name but two.

Thus to conclude, Lawson's account is weak on at least three fronts. First, it does not, in fact, avoid the paradoxes of confirmation since they reappear at a different level, within a field of application. Secondly, the logico-probabilistic approach is not free from counter-intuitive consequences, as he claims (*Ibid.*, p. 400), and the Bayesian way around one of these is not available to it.

Thirdly, and finally, no evidence is given to support the claim that this view concords with observed research practice.

Given these conclusions, one might feel entitled to suggest that perhaps what is needed here is a third approach which goes beyond both the 'logical' and 'historical' attempts. In this respect it would be interesting to see whether some of the more recent discussions of the relationship between theory and evidence, such as Glymour's for example, could usefully be applied in the context of economics theory. It may be true that old dichotomies never die, but perhaps it's long past time for this particular one to fade away and for us to move on to something potentially more fruitful.

STEVEN FRENCH
University of Campinas, Brazil

REFERENCES

GLYMOUR, C. (1980): *Theory and Evidence*. Princeton University Press.

HOWSON, C. (1984): 'Bayesianism and Support by Novel Facts', *British Journal for the Philosophy of Science*, 35, pp. 245–51.

HOWSON, C. (1985): 'Some Recent Objections to the Bayesian Theory of Support', *British Journal for the Philosophy of Science*, 36, pp. 305–9.

LAWSON, T. (1985): 'The Context of Prediction (and the Paradox of Confirmation)', *British Journal for the Philosophy of Science*, 36, pp. 393–407.

Brit. J. Phil. Sci. 39 (1988), 537–540 *Printed in Great Britain*

DISCUSSION

Davy Refuted Lavoisier Not Lakatos

F. Michael Akeroyd has issued a challenge to followers of Lakatos [1986]. He dares them to explain the acceptance of Davy's idea that not all acids are composed of oxygen as Lavoisier had so forcefully claimed. This acceptance supposedly creates a problem for Lakatosians. According to Lakatos:

> For instance there may have been an experiment which was accepted instantly—in the absence of a better theory—as a negative crucial experiment. For the falsificationist such acceptance is part of the internal history; for me it is not rational and has to be explained in terms of external history. (Akeroyd, [1986] p. 360)

Since there was neither real theoretical understanding of acids until much later in the 19th century nor a better theory, Akeroyd throws down the gauntlet: Show me some external factors or show that accepting Davy's results were irrational. Instead of showing either, I shall show that Akeroyd is misguided in his challenge.

First, a brief reminder of the history.

In 1810, Davy was able to show that muriatic acid (hydrochloric acid) contained no oxygen and that chlorine was an element. To re-quote Leicester:

> . . . Davy gave the deathblow to Lavoisier's theory of the composition of acids. (p. 161)

Akeroyd says 'According to the theories of Lavoisier which dominated Chemistry 1790–1810 an acid was the oxide of a non-metallic element.' (Akeroyd [1986], p. 360.). At first blush then, Davy would seem to have overthrown Lavoisier's theories. This account is (1) an oversimplification and (2) irrelevant to Lakatos' point quoted above.

THE OVERSIMPLIFICATION

Lavoisier had no theory about acids. Indeed, it is not at all clear that he had any theories at all. According to both Leicester and Partington, what Lavoisier did (most famously) was to show experimentally that:

(1) the interconvertibility of water and earth was mistaken;
(2) diamonds could be burned in air if enough heat is applied;

(3) gases take part in chemical reactions;
(4) what he terms oxygen was responsible for many of the odd facts which the phlogiston account explained and was itself committed to;
(5) water was a compound of inflammable air (hydrogen) and oxygen.

Leicester cites no evidence that Lavoisier might have had to support the claim that all acids are composed in part of oxygen and neither does Partington [1962]. Leicester says 'Impressed by the importance of oxygen, he believed that all acids contained this principle . . .'. (Leicester [1956) p. 161)

Yet in an important sense it does not matter whether Lavoisier had any, or a great deal of, evidence to support his views on the composition of acids. What is important is that his view on acids was not a theory. It was at best nothing more than an idea based on an experiment. Indeed, in the five discoveries made by Lavoisier cited above, whether taken separately or severally, there is no clear theory to be found nor was there one stated by Lavoisier; at least, not in the sense that Lakatos is using the word, 'theory'.

This is not difficult to show. From Partington, cited approvingly by Akeroyd, we are shown the first announcement of Lavoisier's theory of acids.

> . . . acids are all composed in great part of air, that this substance is common to all and they are differentiated one from another by the addition of different principles particular for each acid . . . Not only air but the purest part of air . . . enters into the composition of all acids without exception, and it is this substance which constitutes their acidity . . . (Partington [1962], p. 414)

Lavoisier then goes on to cite experimental results: what happens when mercury is dissolved in nitric acid. The interpretation of the experiment naturally is in line with the idea that all acids are acids precisely because they contain pure air (oxygen).

Now there are some background beliefs which guided Lavoisier's experimentation and interpretation. These are clear. But what is not clear is: Where is the theory of acids as we and Lakatos would us the term 'theory'? Surely, a hasty generalization, 'All acids contain oxygen' is not a theory.

THE IRRELEVANCY

Lakatos talks about theories in terms of methodologies. It is the methodologies he asks us to evaluate. Better theories for Lakatos are always to be understood in terms of better research programmes. Now we are ready to pick up Akeroyd's gauntlet.

At the period in the history of chemistry upon which Alkeroyd has focussed, there were no competing theories, as Lakatos would have it, because there were no competing research programmes. There were just experimental results. Unless one wants to call good experimental procedure a research programme, there were no research programmes at all. There was no crucial experiment, in this instance, because there was no theory choice to be made. Put another way, Lakatos' views on theory choice apply to mature science, defined by the existence of rival research programmes. At this time, chemistry was not a mature science.

Akeroyd himself comes close to realizing this. He says:

> Berzelius had just finished extending Lavoisier's oxygen theory [*sic*] into a comprehensive 'electro-chemical' or 'dualistic' theory of acids, bases and salts. In the best Lakatosian traditions he was not prepared to give up a promising research programme for the sake of a single empirical result. (Akeroyd [1986], p. 361)

It is precisely because Berzelius had a theory in the Lakatosian sense of research programme that he doggedly kept to the view of Lavoisier. Berzelius' theory did not turn chemistry into a mature science, since the theory was his and not followed by others. To almost all the other researchers, there was no theory and, therefore, the question was one merely of relatively straightforward fact. That is, is there a strong acid which does not contain oxygen? One could hardly fly in the face of Davy's results without an overarching theory.

There was the atomic theory of Dalton published in 1808 in *A New System of Chemical Philosophy*. It is not necessary to quibble over whether Dalton's ideas represent a theory. It need only be noted that at the time of Davy's claim about muriatic acid, the ideas of Dalton were not relevant to the nature and composition of acids. Again, without such a theoretical background, one just cannot apply Lakatos' categories.

In this regard, it is worth noting that the concept of 'single empirical result' is a function of one's philosophy of science—at least as one is attempting to interpret the history of science. To a strict (naive) falsificationist, Davy's result counts as one result to be applied in a *modus tollens* fashion to the rest of the theory. To a Lakatosian, however, if there is a theory involved, an experimental result may be numerically single, but in terms of methodology, has to be evaluated as part of the larger theoretical framework into which it is to be absorbed or not. Whether this is a strength or a weakness of the Lakatosian view is an old question. But in any case, it is Lakatos' view and for this issue that is what is important.

In this light, that Davy's experiment was accepted needs no special interpretation from the followers of Lakatos.

ARTHUR ZUCKER
Department of Philosophy
Ohio University
Athens, Ohio 45701

REFERENCES

Akeroyd, F. M. [1986]: 'A Challenge to the Followers of Lakatos', *The British Journal for Philosophy of Science.*

Leicester, H. M. [1956]: *The Historical Background to Chemistry*. Dover, 1971.

Partington, J. R. [1962]: *A History of Chemistry*. MacMillan, 1962.

Brit. J. Phil. Sci. 39 (1988), 541–550 *Printed in Great Britain*

REVIEW ARTICLE

Cantorian Set Theory and Limitations of Size*

Hallett, Michael. *Cantorian Set Theory and Limitation of Size.* Oxford Logic Guides #10. Oxford: Clarendon Press, 1984. xxii + 343 pp.

Michael Hallett's book is ostensibly a philosophical inquiry into the nature of set theory and the extent to which the axiomatic systems of Zermelo and von Neumann in particular drew their inspiration from the original set theoretic work of Georg Cantor. But its leitmotif, which becomes increasingly apparent in the last third of the book, is a generally negative critique of set theory. Ultimately, Hallett's goal is to demonstrate not only how it is flawed as a theory, but the extent to which, philosophically, it fails to provide a satisfactory foundation for mathematics in general.

Hallett is especially critical of limitation of size theories, which purport to resolve the problem of paradoxes within set theory. To Hallett, these fail to deal satisfactorily with such fundamental concepts as 'set'—and especially power sets. Among other indictments, according to Hallett, is set theory's failure to resolve the Continuum Hypothesis, the Souslin conjecture, and the question of whether all projective sets are Lebesgue measurable. In the end, despite considerable work since Cantor, Hallett suggests that the continuum may not be appropriately reducible to a set theoretic interpretation capable of resolving matters such as these. This would mean that set theory is fundamentally incapable of serving as a foundation for mathematics. But then, as Hallett puts it, 'Why should we expect that all of mathematics should be embraced by it?' (page 305).

Although such a judgment may seem overly pessimistic to some, and certainly premature to others, Hallett argues his case by carefully examining set theory and its development, beginning with the work of Georg Cantor late in the 19th century. Throughout the book he uses historical examples to motivate or illustrate his larger philosophical objectives. Hallett's most important historical claim concerns Cantor's metaphysics, which from the beginning contributed enormously to Cantor's conceptualization of certain key features of set theory. As Hallett emphasizes in his preface, 'there is a direct route from Cantor's metaphysics to the substance and nature of modern set theory,' (page x). Demonstrating this point is one of the main burdens of Hallett's book.

In analyzing the fundamental concept of 'set,' Hallett rightly emphasizes the difficulty of deciding which objects are to be allowed as sets. Cantor's realization that certain collections are too large to be sets did not arise, Hallett says, from his discovery in the 1890's of the paradoxes of set theory stemming from overly large sets. Instead, he argues that the distinction which eventually led to a limitation-of-size theory was a consequence of Cantor's much earlier notion of Absolute infinity, which he had begun to develop for purely philosophical reasons as early as 1882. As Hallett puts it, 'the notion of Absolute infinity was originally presented as a metaphysical doctrine about what is necessarily excluded from mathematical treatment, a doctrine adopted largely for theological reasons, to adapt the new theory of *mathematical* infinity to the traditional Aristotelian-Scholastic teachings on infinity,' (page xiii).

What Hallett does not consider is the possibility that Cantor was already aware of the paradoxes of set theory—at least of the paradox involving the collection of all transfinite ordinal numbers—as early as 1882. In a letter of 1897 Cantor himself told Hilbert that he had expressly introduced the distinction of 'Absolute infinity' in his *Grundlagen* of 1883 because he was aware of the mathematical problems that arose with notions like the 'set of all things.' Even then, however, he seems to have regarded the paradoxes as necessary, natural consequences of set theory rather than as logically disturbing inconsistencies. In the *Grundlagen*, for example, he specifically referred to the 'absolut unendliche Zahlenfolge' and suggested obliquely that the absolutely infinite sequence of numbers was beyond mathematical determination.[1]

On the subject of Cantor's famous transfinite numbers, Hallett goes to great lengths to explain why he should have adopted an ordinal theory of cardinality long before there was any hint of its inevitability. Again, he finds the answer in Cantor's philosophical approach to infinity (pp. 49–118), even though there is a simpler explanation—and again, an historical one.

Cantor's entire theory of infinite sets was motivated by Cantor's study of ordinal numbers, first in the 1870's in terms of indexes for infinite point sets needed for his theorems on trigonometric series. Later these 'indexes' evolved from 'infinite symbols' to 'transfinite ordinal numbers,' and were finally presented as 'ordertypes of well-ordered sets' long before Cantor began to think about transfinite cardinal numbers. Although Cantor's *Grundlagen* of 1883–1884 developed the theory of ordinal numbers in great detail, transfinite cardinal numbers were never mentioned. These eventually appeared a decade later, after having progressively evolved in Cantor's mind from 'powers' of sets to transfinite cardinal *numbers*. Adoption of the transfinite alephs only occurred in the 1890's. By then the entire theory of ordinal numbers was solidly in place, and provided a natural and easy means of defining transfinite cardinal numbers. Rather than adopting an ordinal theory of number for primarily philosophical reasons, it is more likely that Cantor was motivated by purely technical requirements. It seems clear that what he followed was the natural course suggested by the mathematics of set theory as it began to unfold, and this development is reflected in the historical evolution of the theory itself.

---

[1] For recent discussion of Cantor's use of the term 'Absolut' see Walter Purkert and Hans Joachim Ilgauds, *Georg Cantor 1845–1918* (Basel: Birkhäuser, 1987), pages 146–166, especially pp. 150–160.

Hallett devotes an entire chapter to the ordinal theory of powers, and this illustrates well how Cantor's attempt to define ordinal and cardinal numbers as sets was closely related to his view that they arose naturally as abstractions from already given sets (Chapter II, pages 49–118). Hallett is quite right, in fact, to stress that 'a certain conceptual priority has been laid down: numbers derive from sets,' (page 54). But in discussing difficulties with this approach to defining numbers by abstraction (the cardinal numbers or powers were derived by a kind of 'double abstraction' in Cantor's papers of 1895–1897), Hallett might have introduced Frege's scathing critique of Cantor's psychologism. Frege was blunt in showing exactly how unsatisfactory Cantor's approach to explaining the connection between sets and numbers via abstractions was. The difficulties could be overcome, but Cantor's notions of set and number were simply too vague to satisfy a philosopher of Frege's temperament. Here, Hallett's account is particularly successful in explaining why Cantor steadfastly championed his abstractionist view of number and staunchly opposed Frege's approach (page 127).

In examining the development of set theory after Cantor, especially the axiomatizations due to Zermelo and von Neumann, Hallett concentrates on the significance of limitation-of-size arguments used to eliminate the paradoxes from set theory. He is especially interested in showing how 'Cantorian' these are—meaning the extent to which they relied on ordinal aspects of transfinite set theory or upon arguments involving limitation-of-size.

Such ideas go back to Cantor, who explained his own limitation-of-size solution for resolving the paradoxes of the largest ordinal or cardinal numbers in letters to Hilbert and Dedekind in the late 1890's. (Limitation-of-size was first put forward in print, quite independently of Cantor, enthusiastically by Jourdain and more tentatively by Russell shortly after 1900).

Hallett's views on Cantor's reasons for adopting the limitation-of-size approach are unusual but interesting. It is Hallett's contention that Cantor's early theory of the philosophical Absolute was later transformed into a mathematical theory of 'absolute collections,' a transformation which subsequently gave birth to a notion of 'limitation of size,' (page 164). Although Cantor's early ideas on the Absolute can hardly be called a theory, given the informal and quite inexact terms in which the idea was expressed in the *Grundlagen* and in Cantor's philosophical writings later in the 1880's, the eventual appeal he made to limitation-of-size for mathematical reasons in the

1890's was surely one of theoretical necessity—and not simply a transformation of a previously philosophical or theological exigency.

As Cantor began to work out the purely mathematical consequences of the sequences of transfinite ordinal and cardinal numbers, the latter were considerably sharpened in the 1890's once they were elevated from powers to actual numbers and shortly thereafter enshrined as his famous alephs. By then (early in the 1890's) Cantor was forced to recognize the logical difficulties inherent in the twin concepts of both series—namely those of the transfinite ordinal and cardinal numbers. Regarding them as well-ordered sets, any consideration of the corresponding order types would have clearly forced the paradoxes upon him, not for philosophical or theological reasons, but for purely technical ones.

Similarly, it is more likely that Cantor took the limitation-of-size approach to bar the paradoxes from set theory for purely internal reasons at the mathematical core of set theory, rather than consciously mathematizing his earlier philosophical principle of the Absolute (as words like 'transform' or 'translate,' both of which Hallett uses, suggest). Later, relating the mathematical limitation-of-size theory with his earlier, much vaguer statements about the 'Absolute' in the *Grundlagen* would have served Cantor well, if only to establish that he had known about the Absolute, both theologically and mathematically, all along.

This is certainly plausible, for it seems clear that limitation-of-size was the most direct and obvious way, from a strictly mathematical point of view, to handle the paradoxes of set theory that Cantor had himself discovered. Others also took this approach almost immediately in responding to various paradoxes associated with set theory, and presumably without knowledge of Cantor's letters to Hilbert and Dedekind. They certainly did so without making any of Cantor's philosophical or theological presuppositions. As a result, limitation-of-size arguments designed to bar the paradoxes from set theory were developed quite naturally in variant forms very early by Jourdain, Russell, Zermelo and (later) Mirimanoff (pages 176–194). Discussion of their theories comprises a very informative and interesting part of Hallett's book.

Hallett emphasizes that later set theorists did not preserve the connection between 'absoluteness' and the theological Absolute (i.e. God) that was apparently central to much of Cantor's thinking about set theory. It should be stressed, however, that we know of Cantor's views about such things primarily from correspondence, fleeting references in a few footnotes in his early work, and from material he published in non-mathematical journals. In his most mature presentation of his set theory, the *Beiträge* of 1895–1897, there is no mention of God nor is any appeal made to the Absolute in order to provide metaphysical or philosophical foundations for set theory.

On the other hand, to make Cantor's faith in the Absolute seem less peculiar or idiosyncratic, it should not be forgotten that later philosophers and mathematicians have also, from time-to-time, made reference to ideal agents or absolute powers. Hallett takes up the question of super powers in a lengthy discussion of Kant, the problem of completability, and Kant's first antinomy (pages 223–239). Hallett also discusses Wang's notion of 'ideal overview'—the presumed ability of an ideal agent to 'collect together' or 'comprehend' a collection of objects (page 220). Wang describes this as 'infinite intuition which is an idealization.' More recently, Philip Kitcher has called upon infinite intelligence to perform mathematical feats on behalf of an empiricist foundation of mathematics.[2] None of these, however, seems much different from Cantor's reliance on God to perform similar idealizations to guarantee any transfinite operations. Hallett's objections to ideal agents and the operations they perform stem from his belief that they could never complete an impredicative process (page 236). Even if one had recourse to 'super time,' it would still be impossible, he argues, to construct sets like P(w) (page 236).

One of the major problems, in fact, with limitation-of-size arguments involves the question of power sets, which Hallett describes as the 'bane of the limitation of size argument,' (page 217). Hallett argues that power sets are 'indissolubly linked' to the 'unlimited extent of the universe,' (page 205). This of course is precisely where Hallett believes that limitation-of-size arguments fail to protect set theory from the paradoxes, and are inadequate to explain why axiom systems that rely upon limitation-of-size effectively bar the paradoxes (page 205).

Hallett is especially hard on recent defenses of Zermelo-Fraenkel axioms using variations on the limitation-of-size arguments, notably those given by Shoenfield and Wang (page 214). Shoenfield's justification of the von Neumann cumulative hierarchy, for example, is basically an iterative one, and Hallett is critical of notions like completion and completability as used by Shoenfield. He regards these as an 'intrusion of constructivist terminology,' and argues that the collections Shoenfield employs are only assumed to be completable, but that this is never really established. Hallett also criticizes Shoenfield's system as clumsy because of a hazy notion of precedence among stages (page 219).

Wang's attempts to justify the iterative concept of set along with the Zermelo-Fraenkel axioms are even vaguer than Shoenfield's, according to Hallett, and are susceptible to similar traps. He dismisses Wang's approach because it attempts to explain the power set axiom by reference to concepts that 'do not at all capture, and indeed contradict, its nature,' (page 221). Of all those who have attempted to present an iterative stage theory, Boolos in Hallett's view has been the most successful, largely because of his description of 'what the universe looks like when (or even when) impredicative principles are included,' (page 223).

[2] Kitcher, Philip. *The Nature of Mathematical Knowledge*. New York: Oxford University Press, 1983.

On the subject of constructivist terminology, Hallett is especially critical of Shoenfield's use of 'completable' and Wang's use of 'formation,' not only for their unwarranted constructivist implications, but because such terms are also used so broadly as to make arguments based on them circular. Here he agrees with Charles Parsons that the use of such constructivist terminology is misleading. In fact, he denounces it as 'dangerous' because it obscures the realist nature of the theory in question. The realist view, Hallett explains, gives definitions for an object already taken to exist, whereas the constructivist specifies a process whereby a new object is formed (page 237).

Discussing Zermelo's system (Chapter 7, pages 240–269), Hallett takes the separation axiom as a limitation-of-size principle and asks how Zermelo could have been led to formulate it? How did he decide which axioms to use? Hallett believes that Zermelo simply analyzed his well-ordering theorem, and then took the basic principles used there to comprise his set of axioms. Hallett also maintains that Zermelo's axiomatization of set theory was prompted, ultimately, by doubts raised by his proof of the well-ordering theorem (page 253).

Hallett criticizes Zermelo's work for confusion over definite properties and the proper formulation of exactly what a 'definite property' might be (page 269). Given the importance of this question, Hallett discusses subsequent attempts to resolve this difficulty. Skolem, for example, proposed a first-order language; Fraenkel, on the other hand, introduced a recursive list of terms to replace 'definite properties'; and von Neumann proposed an entirely new axiomatization of set theory.

Von Neumann, in fact, emerges as the real hero of Hallett's story precisely because he adopted the replacement principle in order to give an adequate axiomatized formulation of the limitation-of-size principle. Having emphasized all along the Cantorian theory of ordinals as a criterion for determining the extent to which later theories are 'Cantorian,' Hallett concludes that von Neumann's measures up extremely well. For example, at one point Hallett says that Cantor's assumptions about sets come very close to the axioms behind von Neumann's first proof of the enumeration theorem, which is really a corollary to the theorem that the universe of sets is cardinally equivalent to the class of all ordinals.

Hallett devotes the last chapter of his book to von Neumann's 'reinstatement' of the ordinal theory of size. Zermelo's limitation-of-size was directed towards comprehensiveness, as was Fraenkel's (collections that were 'too comprehensive' were not allowed; attention was not directed to the question of what sets, cardinally, were too large). On the other hand, Hallett describes how von Neumann returned to Cantor's original formulation by employing a restriction based upon a cardinal concept of size. In von Neumann's system what was meant by 'too big' was determined in terms of cardinality (page 271).

Hallett regards von Neumann as having simply discarded Cantor's definitions by abstraction as too vague in favor of purely set theoretic formulations.

He defined an ordinal number (as Cantor had in 1883) as the set of all ordinals that precede it. Von Neumann then introduced the Cantorian alephs in the simplest possible way by defining them as initial ordinals (page 276). In this context, Hallett provides a very interesting discussion of the discovery (usually attributed to Fraenkel and Skolem) of the replacement axiom (page 280). However, Hallett argues that von Neumann should be given credit as the true 'discoverer' of the axiom because he was the only one to understand its actual significance and to work out its consequences (page 282).

By the end of the book, Hallett has succeeded in showing that, while the concept of set was originally taken for granted as a primitive concept, one regarded as elementary and which subsequently came to serve as the very foundation for much of modern mathematics, it is not as simple an idea as its early proponents thought. One continuing dilemma surrounds the question of when a collection constitutes a unity—a oneness. Despite attempts to make such notions precise by means of the various axiom systems developed since the time of Zermelo for set theory, Hallett stresses that we still lack any clear sense of why they work.

In his conclusion, Hallett discusses these difficulties, not 'to devalue set theory, but only to draw attention to its complexity and some of its fascination,' (page 299). On the contrary, however, Hallett disparages set theory for being too vague. 'Where set theory suffers as a foundation framework,' he says, 'is that in general it does not bring this conceptual clarity with it. It is no good, philosophically speaking, reducing to set theory something so basic to human thought as the elementary theory of natural number if you cannot also explain why numbers are sets, and why the set concept is even more fundamental. But the set concept is too unclear for any such explanation to be given,' (page 301).

And yet Cantor does succeed in showing how set theory can account for numbers, both finite and infinite, ordinal and cardinal, and especially how set theory serves thereby to clarify considerably the mathematical understanding of number. Cantor's interpretation of the natural and real numbers from a set theoretic point of view, as Hallett's book makes clear, achieved a kind of set theoretic reductionism that put the latter, in particular, on a firm foundation.

Even so, 'does it not seem in the end,' Hallett asks, 'that we are explaining the relatively well-understood (natural number) with the less well-understood (set)?' (page 301). But is there any reason to believe that because mathematicians have known about numbers for a much longer time than they have known about sets, that sets cannot succeed in sharpening the concept of number (as Cantor believed they did)? The issue is whether or not numbers are better understood when interpreted through or given a foundation via set theory. Whether one chooses the approach taken by Dedekind or Cantor, this certainly seems to be the case. As Dedekind asked in the title of his little masterpiece, *Was sind und was sollen die Zahlen?*. Clearly, Dedekind regarded the set theoretic approach as making precise what mathematicians previously had taken for granted as well-understood. In fact, concepts like real number and continuity were not so well understood that they could be satisfactorily explained apart from the clarity and rigor that set theory permitted. The set concept may still require work, but that does not mean that it has failed in providing a better foundation for numbers or for mathematics as a whole (page 301).

As for limitation-of-size, while it may serve in axiomatic frameworks to bar the paradoxes of set theory generated by considering some collections like the universe of all things as 'too large,' Hallett takes the power-set principle to fall outside the limitation-of-size theory precisely because 'it is strong enough to render the classical theory of the continuum,' (page 304). Instead, he argues that axiomatic systems have been advanced for concepts (namely sets) that we cannot easily understand. Moreover, apart from the demands of a theoretical reduction of set theory, Hallett finds that it is hard to say why many of the key collections isolated by the axioms should be sets (page 304).

One might be pragmatic, satisfied that the theory works, simply taking 'carte blanche and exploiting the unexplained axiomatic set notion for all it is worth,' (page 305). But as Hallett says, the conceptual problems will not go away just because the theory in question is successful, a lesson that 'we should have learnt from the history of physics,' (page 305). Indeed, one needn't go so far afield as physics; *mathematics* teaches the same, as one should have learnt from the history of the calculus.

The conceptual difficulties with set theory are made clear, Hallett suggests, by its attendant theoretical difficulties. Set theory has failed to answer key questions about the continuum: does $2^{\aleph_0}=\aleph_1$? Is the continuum characterized by the Souslin property? Is every projective set Lebesgue measurable? These failures as Hallett sees them of set theory may be due to something very simple—perhaps there is no reason to take the continuum as a set-theoretic structure. If so, then there is no reason to expect set theory to be able to resolve every question that may arise about the continuum. The set concept was always unclear, says Hallett, and the paradoxes made a bad situation worse because they encouraged axiomatizations that were advanced using a concept that was inherently vague. As Hallett concludes, maybe sets are fundamental to abstract thought, but why should we expect that all of mathematics should be explicable through the use of sets? (page 305).

Despite excellent use of historical materials, primarily those in Cantor's *Nachlass* in Göttingen, it should by now be clear that it is not the intent of Hallett's book to present an historical account of Cantor's life and work. This sets his book apart from such recent biographical studies as those of Meschkowski, Grattan-Guinness and myself, or of Kertész and Purkert/Ilgauds, both of which were unavailable to Hallett at the time he was writing.

Nor is Hallett's book an internalist historico-philosophical study in the spirit of Gregory Moore's recent *Zermelo's Axiom of Choice: its origins, development and* *influence* (Berlin: Springer, 1982). Hallett saw Moore's book just as his own was going to press. In a brief mention of Moore's study (p. xvi), Hallett is particularly critical of references to 'implicit' or 'unconscious' uses of the Axiom of Choice before Zermelo's famous proof that every set can be well-ordered (1904). The Axiom of Choice, Hallett says, 'was something quite new with Zermelo, and to talk of implicit uses *before* may obscure some of its radicalness, no doubt unintentionally but in any case wrongly,' (page xvi). This seems ironic criticism in light of Hallett's own references to 'the implicit well-ordering assumption in Cantor's own work,' (page 139, and again on page 144).

Among philosophers sensitive to history, Hallett's book is probably best compared with Philip Kitcher's recent *The Nature of Mathematical Knowledge*. Although Kitcher has an explicit philosophical point to make, namely justification of an empiricist account of mathematics, he also employs many historical examples to help illustrate and ground his arguments.[3]

Hallett's book, on the other hand, seeks to make a philosophical point about modern set theory and 20th-century attempts to develop sound and coherent foundations. Only when it suits his purposes does Hallett call upon historical examples, with the result that such material appears sporadically.

Rather than using its history to show how set theory developed from one stage to another in its evolution, Hallett mostly uses history to justify his philosophical claims. At times, he even overlooks historical connections entirely. Take for example his anachronistic statement that 'the most useful point with which to begin tracing Cantor's line of argument is his stress on science as a "free" conceptual construction,' (page 14). Cantor, who was speaking of mathematics (and not science as a whole)—did not himself take this as a starting point. Quite the opposite, for this important philosophical position on the foundation of mathematics was directly related to Kronecker's opposition to Cantor's theory. In fact, all of Cantor's distinctions between intrasubjective and transient realities, his insistence that objects in mathematics must be taken to exist when they are proven self-consistent, including his attempt to relate the acceptance of his new transfinite numbers to traditional acceptance of the irrational numbers (at one point he even called his transfinite ordinal numbers 'new irrationals'), needs to be seen against the background of his acrimonious disagreement with Leopold Kronecker.[4]

Some readers may find annoying the mediocre editing and relatively large number of typographical and other printing errors. Others will undoubtedly object to certain locutions which strain contemporary English usage, whether British or North American, for example 'speaking Cantorianly' (page 117), 'Berkeleyian (sic) way' (page 14), and 'Zermelian' (page 48). Readers may wonder about the usage of some words, as for example on page 55 where

---

[3] See above, note 2.

[4] See J. W. Dauben, *Georg Cantor. His Mathematics and Philosophy of the Infinite*. Cambridge, Mass.: Harvard University Press, 1979, pages 125–137.

Hallett writes 'but the unity obtained is not just objectual.' Does he mean objective? Is 'objectual' here a technical term with a special philosophical connotation that is being used without explanation? The references to 'C.A. Vatson' (both in the text and bibliography) should be corrected to Claude Alphons *Valson*. In fact, throughout the book Hallett routinely refers to minor mathematicians and other figures who, in the absence of an historical glossary, will be unknown to most readers. For example, how many who are not Cantorian specialists will know who Albert Eulenburg (not Eulenberg, as on page 41), Constantin Gutberlet, Everhard (not Eberhard, page 20) Illigens, Ignatius Jeiler, Benno Kerry or Aloÿs von Schmid are? On page 13, read 'Actus Purissimus' for 'Actus Durissimus.' And finally, one assumes that the '?' in the citation to 'Fraenkel ?1927' on page 203 is a misprint, although the reference to page 118 is wrong; perhaps the '?' was meant as a sign to check the reference before going to print?

Surprisingly absent from the bibliography is the important article by Ivor Grattan-Guinness, 'Towards a Biography of Georg Cantor,' *Annals of Science*, 27 (1971), pp. 345–391. Otherwise, the bibliography is lengthy and useful.

After reading Hallett's book, a number of questions remain unresolved, including several very important philosophical issues that have run as leitmotifs, really, throughout the entire book. Why, for example, do the axiomatizations of set theory cohere as theories? Why do these theories seem to work as well as they do? If one still hopes to find an all-embracing foundation for mathematics, and if Hallett is right that set theory, long taken as the prime candidate, should now be placed out of the running, is there any obvious alternative for its replacement?

Einstein spent the last decades of his life searching for a universal principle for physics. Although he hoped to find the solution in what he called a unified field theory meant to combine gravitation and electrodynamics in a single mathematical theory, he was never able to achieve satisfactorily this life-long ambition.

Are mathematicians similarly doomed to disappointment if they too hope for a single unifying principle to comprehend all of mathematics? The human desire for unity may indeed be a very basic one, and in various guises it seems to pervade nearly all areas of human thought, beginning with the Presocratics. The reduction of complex matters to one simple principle that will prove all-inclusive may simply reflect a deep psychological need. Is it this, perhaps, that is ultimately impossible to fulfill? And if set theory comes closest to having achieved this desire for unity in mathematics, is that perhaps enough?

JOSEPH W. DAUBEN
Department of History,
Herbert H. Lehman College, CUNY and
Ph.D. Program in History, The Graduate Center,
City University of New York

Brit. J. Phil. Sci. 39 (1988), 551–561 *Printed in Great Britain*

REVIEW ARTICLE

Observation and Objectivity

BROWN, HAROLD [1987]: *Observation and Objectivity*. Oxford University Press. *ix*+255 pp.

---

1 *A Causal Theory of Perception*
2 *Scientific Objectivity*
3 *Historicism and Epistemology*
4 *Conclusion*

Ever since Hanson [1958] argued that there is more to seeing than meets the eye, contemporary philosophers of science have debated what else is involved. Hanson himself introduced an influential position in this debate, a position suggested by the following remarks:

> What is it to see boxes, staircases, birds, antelopes, bears, goblets, X-ray tubes? It is (at least) to have knowledge of certain sorts. . . . It is to see that, were certain things done to objects before our eyes, other things would result. . . . To see an X-ray tube is at least to see that, were it dropped on stone, it would smash. . . . Seeing a bird in the sky involves seeing that it will not suddenly do vertical snap rolls; and this is more than marks the retina. ([1958], pp. 20–1; *cf.* Hanson [1969], p. 107)

Hanson's remarks suggest an *epistemic* view of visual perception, a view that makes *knowledge* essential to such perception.

Epistemic accounts of perception have been well represented since the publication of Hanson's view. For example, D. M. Armstrong ([1961], p. 112) has argued that 'perception necessarily involves the acquiring of knowledge of (or inclination to believe in) particular facts about the physical world, by means of our senses'. However, many philosophers who claim that perception is epistemic reject the view that perception essentially involves knowledge. Rather, these philosophers hold that perception essentially involves conceptualization, belief, or an inclination to believe. Recent proponents of variations on the latter view include Pitcher [1971], Gregory [1973], [1974], Craig [1976], Scheffler [1982], Heil [1982], [1983], Runzo [1982], Ben-Zeev [1984], Dancy [1985], Pendlebury [1987], and (apparently) Papineau [1987].

A non-epistemic account of perception denies that perceiving is essentially conceptualization or the acquisition of knowledge or belief. Variations on such an account have been defended, for example, by Dretske [1969], [1979], [1981], Cornman [1975], Jackson [1977], and Fodor [1984]. However, proponents of a non-epistemic account of perception are not united by a positive theme. They share only their rejection of epistemic accounts of perception.

In *Observation and Objectivity*, Harold Brown contends that an adequate account of scientific objectivity must draw on both epistemic and non-epistemic accounts of perception. I shall explain and assess Brown's contention by examining his causal theory of perception, his account of scientific objectivity, and his general epistemological method.

I A CAUSAL THEORY OF PERCEPTION

The main aim of Brown's theory of perception is to explain the role of perception in scientific knowledge (p. 173). Thus, Brown recommends that his theory should be evaluated in terms of its ability to explain scientific observation and the role of such observation in scientific theory choice (pp. 225–6). I have no problem with that recommendation.

Brown's account of epistemic perception differs from Armstrong's. Given Armstrong's view that 'perception is nothing but the acquiring of knowledge of the physical world by means of the senses' ([1961], p. 112), we evidently would have to say that scientists see electrons and neutrinos. But Brown finds this implication unacceptable, on the ground that electrons and neutrinos lack visible properties and thus never 'appear' before us (pp. 83–5, 86). This, of course, raises the question of the conditions for an item's *appearing* before us. Let's ask whether Brown's account of perception answers this basic question.

Brown's account begins with this notion of a *visual field*: one's visual field is the complete set of items that one 'could be visually aware of' when one looks in a given direction at a given time (p. 85). This notion, however, is vague on two counts. It is unclear what sense of 'could' is relevant, and it is unclear what notion of visual *awareness* Brown has in mind. Unfortunately, Brown does not remove either unclarity. The sense of 'could' is important here, since the various construals of 'could' (*e.g.*, as logical, physical, or biological) will give us significantly different notions of a visual field. I suspect Brown's notion of could is biological or physiological, since it is supposed to support the view that perceivers with good peripheral vision have wider fields than people with poor peripheral vision. In any case, let's assume such a notion.

Brown's notion of visual awareness is also important, since different notions will result in different notions of a visual field. If visual awareness requires conceptualization, for example, one's visual field (at least typically) will be more restricted than it would be on the assumption that visual awareness is nonconceptual. For conceptual visual awareness requires a sort of intellectual skill and activity of the perceiver that is not required by nonconceptual visual awareness. Unfortunately, Brown does nothing to clarify his basic notion of visual awareness. One can only guess whether he takes visual awareness to be conceptual, and (if so) in what sense he takes it to be conceptual. Moreover, Brown cannot informatively use his subsequent account of epistemic seeing to elucidate his notion of visual awareness, since his notion of epistemic seeing (as we shall see) presupposes his notion of a visual field and thus his notion of visual awareness. Thus, to avoid conceptual circularity, Brown must provide a conceptually independent account of visual awareness. Further, if an item's visually appearing to one is just one's being visually aware of it, we can infer that Brown's account does not clarify the basic notion of visual appearing.

Brown's rough notion of a visual field contributes to his notion of an item that is *distinguished* in one's visual field. According to the latter notion, one distinguishes an item in one's visual field when one 'picks out' that item as a distinct item, irrespective of its ontological status (pp. 85–6). However, Brown denies that one's distinguishing an item entails one's identifying that item. He claims that 'we can often distinguish items that we cannot in fact identify, for example, an utterly unfamiliar material object standing before us' (p. 86). But one might be puzzled by this claim, since it seems that whenever one distinguishes an item as a distinct item, one thereby identifies it *at least as a distinct item*. This raises the issue of what Brown means by 'identifying' an item.

Brown holds that to identify an item is 'to locate it in a conceptual geography, that is, to integrate it into a body of concepts and beliefs' (p. 87; *cf.* pp. 23, 179). But apparently not just any body of concepts will do, since Brown claims that one cannot identify items for which one has 'no appropriate concept' (p. 87). The latter claim seems trivial if we drop the modifier 'appropriate'; and if we keep that modifier, at least two problems arise. First, we are left with the unanswered, difficult issue of the exact conditions for a concept's being appropriate in the relevant sense. (Brown (pp. 119, 176–9) plausibly rejects the view that 'appropriate' means 'correct'.) Second, given Brown's distinction between distinguishing and identifying, we are faced with the question why one's identifying an item can not involve one's identifying it just as a distinct item. That is, we need to know exactly how one's identifying an item differs from one's distinguishing it. It is not very helpful here to say that one's identifying an item is one's establishing 'its possible relations to other kinds of entities' (p. 23), since it is unclear how such establishing goes beyond distinguishing. A better alternative to Brown's distinction claims that distinguishing is at least a minimal form of identifying, and that some forms of identifying are conceptually more complex than others. I fail to see why Brown does not endorse this plausible alternative.

In any case, we now can introduce Brown's characterization of epistemic seeing: To epistemically see an item is to distinguish it in a visual field and to identify it in terms of an available conceptual framework (p. 86). There are two noteworthy consequences of this characterization. First, what one epistemically sees depends not only on one's visual field, but also on one's set of concepts. Thus, two people with relevantly different conceptual systems can epistemically see different items when they look at a single object (p. 88). For example, as Hanson [1958] notes, Kepler and Tycho saw different things when they looked at the sun. Second, one's epistemically seeing an item does not entail that it exists in the physical world (pp. 119–21). For example, medieval astronomers epistemically saw a sun that moves around the earth; but this sun does not exist.

The latter claim is plausible, of course, only if it is understood as the claim that the sun does not exist *as the sort of object* characterized by medieval astronomers. Thus, we apparently should understand Brown's view to be that one's epistemically seeing *X* does not entail that *X* exists *as the sort of object* conceptualized by one. However, Brown's view does imply that 'a physical object must act on my senses and produce an appropriate effect in me if I am to perceive' (p. 149). Following Churchland [1979], Brown takes such causal interaction of a visual sort to be *non*-epistemic seeing (pp. 80–1, 197–8). Thus, Brown's account allows us to say that Kepler and Tycho non-epistemically saw the same sun, and that the sun non-epistemically seen by the medievals does exist. However, as anticipated above, it is not clear what sort of visual awareness Brown requires for non-epistemic seeing. In fact, it is not even clear that Brown requires a *psychological* state beyond the physiological state of the determinate stimulation of one's visual receptors (p. 86). I suggest, in any case, that there is a relevant difference between one's merely looking at an object and one's non-epistemically seeing it; the latter, unlike the former, requires a sort of visual *awareness*, such as visual attention-attraction (which should not be confused with selective attention-focusing). It is unclear how Brown's account would handle this relevant difference. (I have tried to explain the epistemic significance of attention-attraction in my [1985] and [1989].)

One of Brown's main concerns is to explain how epistemic seeing can be a source of information about items we cannot see, such as electrons and neutrinos. This concern relies on the following notion of observation:

> To observe visually an item, *I*, is to gain information about *I* from the [visual] examination of another item, *I**, where *I** is an item that we epistemically see, and *I* is a member of a causal chain that produced *I**. (p. 93)

Brown qualifies this notion with two points: first, to observe *I* via *I** one must have 'sufficient background information to establish the causal connection between *I* and *I**'; and second, to claim that one is observing *I* via *I** is to claim that *I* is a necessary component of 'the only acceptable [=reasonable] causal explanation' of *I** (p. 93). Even if the latter two points can handle the familiar problem of deviant causal chains (on which see Peacocke [1979])—and it is not obvious that they can, one might wonder whether the second point is too demanding. It seems that one might claim to observe *I* via *I** even while granting that *I* is an essential part of just one of two equally superior causal explanations of *I**. That is, claims to observation seem not to be precluded just by one's acknowledging the existence of two equally superior, empirically equivalent explanatory systems. Or, at least, the contrary view demands considerable argument.

Brown holds that we *observe* physical objects by epistemically perceiving other, intermediary items, *viz.*, subjective percepts. He denies that we can epistemically perceive items in the physical world (pp. 175, 184; *cf.* pp. 149, 157, 187). Thus, regarding epistemic perception of the external world, Brown rejects direct realism. His causal theory is a representational theory in the general spirit of Locke's, although it aims to be compatible with a materialistic ontology (pp. 153, 156).

However, Brown has not provided the needed argument for his claim that 'we cannot perceive the external world' (p. 184). Nor does his theory require this claim. Nothing in his notion of epistemic seeing rules out the possibility that there is an item in the external world that I have distinguished in my visual field, and have identified via my conceptual framework. Brown has given us no reason to think that one cannot be visually aware of an item in the external world, and thus he has given us no reason to think that one's visual field cannot include an item in the external world. Further, he has given us no reason to think that one cannot distinguish and identify an item in the external world. The needed reason is not provided by the claim that 'the items we epistemically perceive are "concept-laden", and are thus mental at least to the extent that concepts are mental' (p. 149). For even if epistemic perception is essentially conceptual, it does not follow that an *object* of epistemic perception is essentially conceptual, or that we epistemically perceive concepts. (Analogously, believing is essentially dispositional, but it does not follow that an object of believing, such as a statement, is essentially dispositional, or that we believe dispositions.) Moreover, the needed reason is not provided by the claim that the items we epistemically perceive 'may not exist independently of our perception of them', or even by the claim that 'these items do not exist apart from our perception of them' (p. 185). We can grant Brown the latter claims for the sake of argument, and then ask how they justify his claim that we *cannot* epistemically perceive the external world. Clearly, the needed justification is not thereby forthcoming.

Now since Brown's notion of epistemic seeing does not rule out epistemic perception of the external world, we can see that it is compatible with direct perceptual realism. That is, contrary to the suggestion of page 157, his notion does not entail the existence of perceived intermediaries between the perceiver and the external world. Rather, his notion entails only that the objects of perception (external or internal) have been distinguished in a visual field and identified. This means that Brown's notion of epistemic seeing does not logically commit one to a representational, indirect realist account of perception. Or, more generally, even a proponent of an epistemic account of perception can consistently endorse direct realism. For an epistemic component in the *process* of perception does not preclude the mind-independence of the *object* perceived. Thus, from the standpoint of the notion of epistemic seeing, Brown's commitment to a Lockean representational account is optional. And I am not convinced that this option is really superior to an adverbial version of direct realism as developed, for example, by Chisholm [1957], Cornman [1975], and Dicker [1980].

Before turning to the topic of scientific objectivity, we should note that Brown's account includes an epistemic notion of *veridical* perception. We might think that one's perceiving an object is veridical if and only if one perceives the object as it actually is. However, Brown claims that the notions of veridical and nonveridical perception 'only make sense in an epistemic context' (p. 174). His support for this claim is thin and unclear, but it involves the additional claim that one will consider one's perception of an object to be veridical if (and, presumably, only if) one has not found that one has misidentified the object. Brown claims that one 'consequence' of the latter claim is that two perceivers can observe the same physical object while epistemically perceiving incompatible items, even though each perceives veridically (p. 175). But this is not really a consequence at all. For conditions for one's merely *considering* perception to be veridical do not entail conditions for perception's actually being veridical. (Analogously, conditions for one's merely believing a proposition to be true do not entail conditions for a proposition's actually being true.)

Nonetheless, Brown's notion of veridical perception gains some plausibility from its being contrasted with the notion of *correct* perception (p. 175). As Brown understands veridical perception, it is just one's conceptually identifying a perceptual object in conformity with the criteria provided by one's conceptual system. Thus, unveridical perception is one's identifying a perceptual object via a concept to which it does not conform (p. 176). Perhaps this is part of a plausible *epistemic* approach to veridical perception, providing a notion of *conceptually coherent* perceptual identification. But it lends no credibility to Brown's aforementioned claim that the notion of veridical perception makes sense *only* in an epistemic context. For, so far as Brown's account goes, we can still talk intelligibly about one's perceiving an object *as it actually is*, independently of epistemic considerations.

Let us turn now to the role of non-epistemic perception in Brown's account of scientific objectivity.

2 SCIENTIFIC OBJECTIVITY

A central aim of Brown's theory of perception is to explain how scientific observation provides for the objectivity of scientific knowledge. Two necessary conditions of scientific objectivity, on Brown's account, are relevance and independence. The requirement of relevance demands that an objective claim be tested, directly or indirectly, against the item about which the claim is made (p. 191). And the requirement of independence requires that an objective claim be about items, or the effects thereof, that exist and have their properties independently of that claim (p. 193). This latter requirement entails a role for non-epistemic perception in scientific objectivity. As Brown sees it, 'scientific observation is always aimed at making sense of information that exists independently of our beliefs', and this non-epistemic information is the input from non-epistemic perception (p. 197).

Non-epistemic information, on Brown's account, constrains, but does not dictate, the possible interpretations, or identifications, involved in the observation of an item. Such information is compatible with a range of interpretations, but it is also incompatible with numerous interpretations. According to Brown, the latter point 'captures the way in which the physical world constrains our observations, and it is through this constraint that observation provides the foundations for objectivity' (p. 199).

Brown thus suggests that non-epistemic perceptual information is incompatible with various perceptual and observational interpretations. But this is a potentially misleading suggestion; for it actually depends on Brown's view that '*given a specific conceptual scheme*, there will be many ways in which a particular item cannot be misidentified', if there is to be genuine perception or misperception rather than fantasy or free association (p. 179). By way of analogy, if I follow ordinary English, I cannot read Brown's *Observation and Objectivity* as Popper's *Conjectures and Refutations.*

Thus, Brown's view is actually that the constraint in question is provided by non-epistemic perception *plus* a specific conceptual scheme. For the relevant thesis is that *given a specific conceptual scheme,* we get obvious incoherence if we interpret non-epistemic perceptual information in certain ways (If this is not Brown's thesis, I fail to see how the constraint in question is epistemically relevant in the sense of pp. 82–3, *i.e.*, in the sense that it enables scientists to decide what they have learned via perception.) We might call this a scheme-relative negative constraint, since it eliminates certain interpretations relative to a conceptual scheme. However, one basic question about objectivity remains unanswered: why, from the standpoint of scientific objectivity, should we prefer one specific conceptual scheme, such as our present scheme, over others. Or, to raise a more serious challenge, why cannot we always satisfy the constraint in question just by loosening up our conceptual scheme in *ad hoc* ways, for example, by including Goodman-style concepts such as grue and bleen. If this is possible, Brown's constraint loses its epistemic teeth. What we need, of course, is a means of testing the relative epistemic merits of alternative conceptual schemes. Given that Brown's constraint is scheme-relative, I fail to see how it can serve this purpose.

In any case, Brown does not take the condition of relevance and independence to be sufficient for scientific objectivity. He holds that for a claim to be objective, it not only must be about some relevant, independent aspect of the world, but also must be capable of being tested and retested in new ways against that aspect or its effects (pp. 202–3). Accordingly, Brown adds that for one to accept a claim in an objective manner, one must be prepared to carry out new tests when they become available. This seeems plausible enough, but I doubt that it identifies what is really distinctive of one's accepting a claim in an objective manner: how one conducts one's believing in light of the relevant tests. Willingness to carry out relevant tests is one thing, while *proper reaction* to the tests is quite another. The former, of course, does not entail the latter; and the latter, for all its complexity, seems to be what is distinctive about the objective acceptance of a claim. Thus, I am not convinced that Brown has identified the hallmark of scientific objectivity, even if he has isolated some necessary conditions.

3 HISTORICISM AND EPISTEMOLOGY

On Brown's view, scientific knowledge consists of scientific claims that have been accepted in an objective manner, *i.e.*, claims that have 'objectively warranted assertibility', (p. 222). Brown thus proposes that we revise our concept of knowledge to sever its definitional tie to the concept of truth. Let us ask what sort of epistemological method underlies this proposal.

Brown's epistemological method seems to be a version of historicism in the sense that it assumes that epistemic norms and principles must be justified by an examination of the history of science, including contemporary science. Thus, Brown claims:

> Just as claims that we make about items in a particular scientific domain must be tested against those items, so claims we make about the methods of acquiring knowledge must be tested by an examination of human experience in the search for knowledge. There are two reasons why we should take the sciences as the basis for any such study. First, they provide a relatively clear example of a persistent and systematic search for knowledge. Second, in the contemporary world an epistemology that cannot make sense of science is *ipso facto* unacceptable, so we might as well begin with science. (p. 32; *cf.* pp. 226–7)

Brown adds that he still thinks there is room for critical examination of scientific methods. The point he wants to stress, however, is that epistemology must take the sciences such as physics, biology, and sociology as its domain, and attempt to construct theories that (a) function analogously to basic scientific theories and (b) are testable in the general way those basic theories are testable (p. 33). Whereas physics, for example, must be testable via various aspects of the physical world, an epistemology must be testable via various scientific theories such as those of physics, biology, and sociology. On this view, epistemology is a second-order science.

I have two main problems with Brown's historicism. First, it is not clear how it actually leaves room for (epistemically) critical examination of scientific methods. If such methods are the ultimate basis for the testing of all epistemic norms and principles, then there evidently will be no further basis on which those methods themselves can be critically examined. If there is such a further basis, then apparently the scientific methods in question will not really be the ultimate basis for all testing. However, one might propose that certain scientific methods can be critically evaluated relative to *other* scientific methods. This proposal leads to my second problem.

The second problem is that the history of scientific methods, norms, and practices is highly diversified in a way that makes any *unqualified* appeal to 'the sciences' problematic. (This point receives implicit support in Toulmin [1972] and in Laudan [1977].) Brown is fond of saying that physics, or the physical sciences, must decide whether certain perceived items actually exist (pp. 122, 127). And historicism makes a similar appeal to the sciences for the justification of epistemic norms and principles. But this raises the question of *whose* construal, or *whose* version, of the physical sciences is relevant to these matters. Can one reasonably appeal, for example, to Lysenkoism in earlier Soviet biology to settle questions about modern genetics, or to Velikovsky to settle questions about the origin of various planets? Surely not, according to many scientists and philosophers.

Brown's more informative thesis is that we must appeal to 'the best available physical theories' (pp. 185, 189). Evidently, this appeal requires a departure from thoroughgoing historicism. Talk of *the best* available sciences presupposes a standard of evaluation for the sciences themselves. And that standard, if we are to avoid vicious circularity, apparently cannot be the result just of an appeal to what *the sciences* recommend or exemplify. Further, given the considerable diversity of science as a social process, not all scientific theses, values, and practices can consistently be treated as epistemically equal. Brown's appeal to 'the best available sciences' acknowledges this. Thus, even on Brown's historicism we need epistemic norms that are not provided just by an appeal to what *the sciences* recommend or exemplify. Consequently, thoroughgoing historicism goes by the board. The actual sciences cannot be our *only* test for epistemic norms and principles. The foregoing quotation from Brown obscures this important point.

What, then, provides the epistemic basis for our assessments of the sciences? Without pretending to give a complete answer, I want to recommend that such a basis can be found via analogy to Goodman's method of reflective equilibrium. Goodman [1979] proposes the following method for the justification of deductive and inductive inferences and rules:

> Rules and particular inferences alike are justified by being brought into agreement with each other. A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend. The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either (p. 64)

I suspect that a similar process of mutual adjustment underlies the justification of epistemic norms and principles. We have considered judgments about the justification of particular non-epistemic propositions, and we have considered judgments about various epistemic norms and principles. By bringing the two sets of judgments into agreement, we reach a sort of explanatory coherence that provides justification for our epistemic norms and principles, and thereby gives us a basis for assessing the sciences. On this approach, we can give a central epistemic role to certain methods and norms in the sciences, but we need not endorse thoroughgoing historicism. The rest of this important story, however, must be left for now as unfinished business. (I say more about it in my [1989].)

4 CONCLUSION

In sum, then, I have raised several notable problem-questions for Brown's account of observation and objectivity. Nonetheless, I want to emphasize that Brown's book is a useful contribution to the literature on scientific observation. It is replete with philosophically important examples of actual scientific observation, and in that respect is a significant improvement on many philosophical works on perception. It seems safe to say that Brown's book is useful reading for anyone interested in the nature of scientific observation.

PAUL K. MOSER
Loyola University of Chicago

REFERENCES

Armstrong, D. M. [1961]: *Perception and the Physical World.* Routledge & Kegan Paul.

Ben-Zeev, A. [1984]: 'The Passivity Assumption of the Sensation-Perception Distinction', *The British Journal for the Philosophy of Science*, 35, pp. 327–43.

Chisholm, R. M. [1957]: *Perceiving: A Philosophical Study.* Cornell University Press.

Churchland, P. M. [1979]: *Scientific Realism and the Plasticity of Mind.* Cambridge University Press.

Cornman, J. W. [1975]: *Perception, Common Sense, and Science.* Yale University Press.

CRAIG, E. J. [1976]: 'Sensory Experience and the Foundations of Knowledge', *Synthese* 33, pp. 1–24.

DANCY, J. [1985]: *An Introduction to Contemporary Epistemology*. Basil Blackwell.

DICKER, G. [1980]: *Perceptual Knowledge*. D. Reidel.

DRETSKE, F. I. [1969]: *Seeing and Knowing*. University of Chicago Press.

DRETSKE, F. I. [1979]: 'Simple Seeing', in D. F. Gustafson and B. L. Tapscott (*eds*), *Body, Mind and Method*, pp. 1–15. D. Reidel.

DRETSKE, F. I. [1981]: *Knowledge and the Flow of Information*. M.I.T. Press.

FODOR, J. [1984]: 'Observation Reconsidered', *Philosophy of Science* 51, pp. 23–43.

GOODMAN, N. [1979]: *Fact, Fiction, and Forecast*, 3rd ed. Hackett Publishing Company.

GREGORY, R. L. [1973]: *Eye and Brain: The Psychology of Seeing*, 2nd ed. McGraw-Hill.

GREGORY, R. L. [1974]: 'Perceptions as Hypotheses', in S. C. Brown (*ed*.), *Philosophy of Psychology*, pp. 195–210. Macmillan.

Hanson, N. R. [1958]: *Patterns of Discovery*. Cambridge University Press.

HANSON, N. R. [1969]: *Perception and Discovery*. Freeman, Cooper and Company.

HEIL, J. [1982]: 'Seeing is Believing', *American Philosophical Quarterly*, 19, pp. 229–40.

HEIL, J. [1983]: *Perception and Cognition*. University of California Press.

JACKSON, F. [1977]: *Perception: A Representative Theory*. Cambridge University Press.

LAUDAN, L. [1977]: *Progress and Its Problems*. University of California Press.

MOSER, P. K. [1985]: *Empirical Justification*. D. Reidel.

MOSER, P. K. [1989]: *Knowledge and Evidence*. Forthcoming, Cambridge University Press.

PAPINEAU, D. [1987]: *Reality and Representation*. Basil Blackwell.

PEACOCKE, C. [1979]: *Holistic Explanation*. Clarendon Press.

PENDLEBURY, M. [1987]: 'Perceptual Representation', *Proceedings of the Aristotelian Society*, 87, pp. 91–106.

PITCHER, G. [1971]: *A Theory of Perception*. Princeton University Press.

RUNZO, J. [1982]: 'The Radical Conceptualization of Perceptual Experience', *American Philosophical Quarterly*, 19, pp. 205–17.

SCHEFFLER, I. [1982]: *Science and Subjectivity*, 2nd ed. Hackett Publishing Company.

TOULMIN, S. [1972]: *Human Understanding*. Princeton University Press.

THE BRITISH SOCIETY FOR THE PHILOSOPHY OF SCIENCE

Programme of Meetings 1988–9

Unless otherwise announced, meetings are held in the Hall of Dr Williams Trust at 14 Gordon Square, London WC1 0AG at 5.15pm on the following dates. Tea and biscuits will be served at 5.00pm.

1989

January 9th	George Ross *The Blue Guitar: Inverse Optics in a New Light*
February 6th	Colin Howson *Prediction and Accommodation*
March 6th	Nancy Cartwright *Abstraction and Explanation*
April 24th	David Gooding *The Experimenter's Redress*
June 5th	4.00pm **Annual General Meeting** 4.30pm Tea and Biscuits John Lucas *Title to be announced*

Nominations for the Elections of Officers and Ordinary Members of the Committee for the year 1989–90 should reach the Honorary Secretary by May 4th 1989.

September	Annual Conference at the University of Reading: *details to be announced*

Philosophy and Computing Science

A Special Interest Group of the British Society for the Philosophy of Science

---

The British Society for the Philosophy of Science Special Interest Group in Philosophy and Computing Science is pleased to announce the following programme of seminars for the Spring Term of 1989.

The seminars will be held in the **Lecture Room, Linacre College, Oxford,** beginning at 5.00pm on the following dates.

All are welcome to attend.

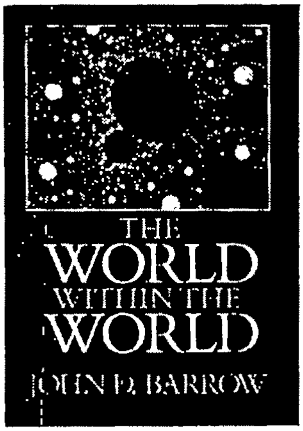
January 23rd 1989	**Dov Gabbay** Imperial College, London **What is a reasoning system?**
February 20th 1989	**Bill Sharpe** Hewlett-Packard Laboratories, Bristol **Knowledge Engineering** **—tasks and domains**
March 20th 1989	**Ned Block** MIT **Title to be announced**

The meetings are to be held in conjunction with Linacre College, and are sponsored by **Michael Jackson Systems Ltd.**

NEW FROM OXFORD

The World Within the World

John D. Barrow



Do there really exist laws of Nature?

Do they have an objective reality, or have we just invented them to make sense of our experience?

Why are they so well described by mathematics?

Is it even possible that there really are no laws of Nature—that they are an illusion?

How does our own existence limit what can be known about the Universe?

This book presents a wide-ranging interdisciplinary study of the evolving concept of laws of Nature. From the magical notions of primitive cultures to the latest ideas about chaos, black holes, inflation, and superstrings the author traces the gradual development of our concept of what laws of Nature are and how we come to know them. The scientific and mathematical topics discussed are of major importance to philosophers and scientists. This book gives them the opportunity to see these new ideas discussed in a serious but non-technical style within a historical context.

0 19 851979 6, 420 pp., illus., Clarendon Press,
May 1988 £20.00

New in Paperback

The Anthropic Cosmological Principle

John D. Barrow and Frank J. Tipler

This book addresses a central problem for both scientists and philosophers: the position and role of man in the universe.

'a remarkable book and a masterly exposition of . . . one of the most important developments to have taken place in physical science.' *Times Literary Supplement*

'an engaging book . . . will be much quoted, much debated, and much praised.' *Nature*

Oxford Paperbacks
0 19 282147 4, 640 pp., illus., April 1988 £9.95
0 19 851949 4, hardback, 1986 £27.50

OXFORD UNIVERSITY PRESS